

Distribution-Free Detection of Structured Anomalies: Permutation and Rank-Based Scans

Ery Arias-Castro¹, Rui M. Castro², Ervin Tánčzos², and Meng Wang¹

¹University of California, San Diego

²Technische Universiteit Eindhoven

Abstract

The scan statistic is by far the most popular method for anomaly detection, being popular in syndromic surveillance, signal and image processing, and target detection based on sensor networks, among other applications. The use of the scan statistics in such settings yields a hypothesis testing procedure, where the null hypothesis corresponds to the absence of anomalous behavior. If the null distribution is known, then calibration of a scan-based test is relatively easy, as it can be done by Monte Carlo simulation. When the null distribution is unknown, it is less straightforward. We investigate two procedures. The first one is a calibration by permutation and the other is a rank-based scan test, which is distribution-free and less sensitive to outliers. Furthermore, the rank scan test requires only a one-time calibration for a given data size making it computationally much more appealing. In both cases, we quantify the performance loss with respect to an oracle scan test that knows the null distribution. We show that using one of these calibration procedures results in only a very small loss of power in the context of a natural exponential family. This includes the classical normal location model, popular in signal processing, and the Poisson model, popular in syndromic surveillance. We perform numerical experiments on simulated data further supporting our theory and also on a real dataset from genomics.

1 Introduction

Signal detection (and localization) is important in a large variety of applications, encompassing any situation where the goal is to discover patterns or detect/locate anomalies. Our focus is on the detection of anomalous behavior which is endowed with some structure. For instance, one might have data consisting of the physical location of a sensor and the corresponding measurement, and would like to determine if there is a spatial region where measurements are unusually high (Balakrishnan and Koutras, 2002). A standard way to tackle this problem is the use of a scan statistic which essentially inspects all (or at least a large number of) possible anomalous patterns. It usually corresponds to a form of generalized likelihood ratio test (Kulldorff, 1997). In (Cheung et al., 2013) the scan statistic was used to detect small geographic areas with large suicide rates and (Guerriero et al., 2009) used the scan statistic for target detection using distributed sensors in a two dimensional region. Although computationally this approach might be challenging, there are a number of situations where it is possible to compute the scan statistic in nearly linear time (Arias-Castro et al., 2005; Neill, 2012; Neill and Moore, 2004; Walther, 2010).

For the purpose of illustration, consider the following prototypical example¹: suppose we have event data over a certain time period and want to detect if there is a time interval with an unusually

¹In fact, this setting might have been the original motivation for the work on the scan statistic (Wallenstein, 2009).

high concentration of events. To make things more concrete and move towards the setting we consider in this paper, assume one can model these event data as a realization of a Poisson process and bin the data, so that we observe a sequence of Poisson random variables. The scan statistic in this particular case combines sums of these values over (discrete) intervals of different sizes and location, together with some normalization — see (2) further down. In this scenario we want to perform a hypotheses test, where the null hypothesis is that no anomaly is present (a homogenous Poisson process) versus the alternative where some intervals have an elevated rate of events (an inhomogenous process). If the (constant) rate is known under the null, then the null distribution is completely specified and the test can be calibrated either analytically or by Monte Carlo simulation. But what if the null event rate is unknown? What are possible ways to properly calibrate the test? What is the price to pay in terms of power?

One can regard the scan statistic as a comparison between observations in one interval to those outside the interval. This point of view leads naturally to a two-sample problem for each interval, which is then followed by some form of multiple testing since we scan many intervals. Thus drawing from the classical literature on the two-sample problem, two approaches can be considered:

- *Calibration by permutation.* This amounts to using the permutation distribution of the scan statistic for inference (detection/estimation).
- *Scanning the ranks.* This amounts to replacing each observation with its rank before scanning. Calibration of such a test can be done by Monte Carlo simulation before the observation of data, as long as the size of the data is known.

The perspective offered by the two-sample testing framework makes these two procedures very natural. The permutation scan has been suggested in a number of papers and applied in a number of ways in different contexts. It is a standard approach in neuroimaging (Nichols and Holmes, 2002) and is suggested in syndromic surveillance (Huang et al., 2007; Kulldorff et al., 2005, 2009). It was suggested by Walther (2010) in the context of a sensor network with binary output and by Flenner and Hewer (2011) in the context of detecting a change in a sequence of images.

Surprisingly enough, the method based on ranks appears to be relatively new in the present context. It was specifically (and simultaneously) proposed as a standalone procedure by Jung and Cho (2015)², where the authors compute the scan statistic on ranks instead of the data itself. Nevertheless, rank-based methodologies have been used earlier in similar settings, but with a different purpose in mind. For instance, the use of ranks in the context of the scan statistic also appears in (McFowland et al., 2013) through the computation of empirical P-values. It is important to note that the use of ranks in the last reference is of a rather different nature than that we propose in our work, and that the emphasis in that paper is on the ability to efficiently compute/approximate scan statistics, while in our work the emphasis is on the calibration of scan tests when the null distribution is not known.

Although less popular, as in the two-sample testing setting, a procedure based on ranks offers some significant advantages over calibration by permutation: (i) it is more robust to outliers and ; (ii) its calibration can be done by Monte Carlo simulation and requires only the knowledge of the sample size³. Point (ii) is rather pertinent, as computationally this is a huge advantage over calibration by permutation. Furthermore, this property is rather advantageous if one desires to apply

²This article was made public after our paper was posted on the arxiv.org. To the best of our knowledge, this other publication became publicly available on October 20, 2015 (doi:10.1186/s12942-015-0024-6), a couple of months after ours appeared online on August 12, 2015.

³The latter explains why, in two-sample testing, methods based on ranks were feasible decades before methods based on permutations, which typically require access to a computer.

the test repeatedly on several datasets of same size; compare with a calibration by permutation: typically, several hundred permutations are sampled at random and, for each one of them, the scan statistic is computed, and all this is done each time the test is applied.

In this paper we study the performance of both the permutation and rank scan methods, providing strong asymptotic guarantees as well as insights on their finite-sample performance in some numerical experiments. In the context of a natural exponential family — which includes the classical normal location model and the Poisson example above — we find that the permutation scan test and the rank scan test come very close to performing as well as the oracle scan test, which we define as the scan test calibrated by Monte Carlo with (clairvoyant) knowledge of the null distribution. We perform numerical experiments on simulated data which confirm our theory, and also some experiments using a real dataset from genomics.

As specified below, we focus on a “static” setting, where the length of the signal being monitored is fixed a priori. Adding time is typically done by adding one ‘dimension’ to the framework, as done for example in (Kulldorff et al., 2005).

1.1 General setting

A typical framework for static anomaly detection — which includes detection in digital signals and images, sensor networks, biological data, and more — may be described in general terms as follows. We observe a set of independent random variables, denoted $(X_v : v \in \mathcal{V})$, where \mathcal{V} is a finite index set of size N . This is a snapshot of the state of the environment, where each element of \mathcal{V} corresponds to an element of the environment (e.g., these correspond to nodes of a network, pixels in an image, genes, etc.). In this work we take a hypothesis testing point of view. Under the null hypothesis, corresponding to the nominal state when no anomalies are present, these random variables are Independent and Identically Distributed (IID) with distribution F_0 . Under the alternative, some of these random variables will have a different distribution. Formally, let $\mathbb{S} \subset 2^{\mathcal{V}}$ denote a class of possibly anomalous subsets, corresponding to the anomalous patterns we expect to encounter (this would be a class of intervals in the example that we used earlier). Under the alternative hypothesis there is a subset $\mathcal{S} \in \mathbb{S}$ such that, for each $v \in \mathcal{S}$, $X_v \sim F_v$ for some distributions $F_v \neq F_0$, and independent of $(X_v : v \in \mathcal{V} \setminus \mathcal{S})$, which are still IID with distribution F_0 . In a number of important applications the variables are real-valued and the anomalous variables take larger-than-usual values, which can be formalized by the assumption that each F_v stochastically dominates⁴ F_0 . We take this to be the case throughout most of the paper. While the standard scan test is calibrated by Monte Carlo by repeated sampling from the null distribution F_0 , in contrast, the procedures we study here — the permutation scan test and the rank scan test — are calibrated without any knowledge of F_0 and F_v .

1.2 Exponential models

Although some of our results will be presented in the general setting above, it is useful to consider an important special case. This serves as a benchmark we can use to compare the performance of the proposed procedures against that of the optimal tests. Doing so is classical in the literature on nonparametric tests (Hettmansperger, 1984), where such a test is compared with the likelihood ratio test in some parametric model (often a location model or a scale model).

In this paper we consider a generic one-parameter exponential model in natural form. Let F_0 be a probability distribution on the real line with all the moments finite. This distribution can be

⁴For two distribution (functions) on the real line, F and G , we say that G stochastically dominates F if $G(t) \leq F(t)$ for all $t \in \mathbb{R}$. We denote this by $G \geq F$.

either continuous (i.e., diffuse), discrete (i.e., with discrete support) or a mixture of both. In the exponential model there is a parameter θ_v , associated with each $v \in \mathcal{V}$, and the distribution $F_v \equiv F_{\theta_v}$ is defined through its density f_{θ_v} with respect to F_0 : for $\theta \in [0, \theta_*)$, define $f_\theta(x) = \exp(\theta x - \log \varphi_0(\theta))$, where $\varphi_0(\theta) = \int e^{\theta x} dF_0(x)$ and $\theta_* = \sup\{\theta > 0 : \varphi_0(\theta) < \infty\}$, assumed to be strictly positive (and possibly infinite). In other words, f_{θ_v} denotes the Radon-Nykodym derivative of F_{θ_v} with respect to F_0 . Since a natural exponential family has the monotone likelihood ratio property⁵, it follows that F_θ is stochastically increasing in θ (Lehmann and Romano, 2005, Lem 3.4.2). In particular, we do have $F_\theta \geq F_0$ for all $\theta > 0$. Important special cases of such an exponential model include the normal location model — with F_θ corresponding to $\mathcal{N}(\theta, 1)$ — standard in many signal and image processing applications; the Poisson model — with F_θ corresponding a Poisson distribution — popular in syndromic surveillance (Kulldorff et al., 2005); and the Bernoulli model (Walther, 2010) with F_θ corresponding to a Bernoulli distribution.

Note that in the formulation above the alternative hypothesis is composite. Tackling this problem using a generalized likelihood ratio approach is popular in practice (Kulldorff, 1997) and often referred to as the scan test, as it works by scanning over the possible anomalous sets to determine if there is such a set that is able to “explain” the observed data. Assuming the nonzero θ_v ’s are all equal to θ under the alternative, and that all subsets in the class \mathbb{S} have same size, some simplifications lead to considering the test that rejects for large values of the scan statistic

$$\max_{\mathbb{S} \in \mathbb{S}} \sum_{v \in \mathbb{S}} X_v. \quad (1)$$

When the subsets in the class \mathbb{S} may have different sizes, a more reasonable approach includes a normalization of the partial sums above, leading to the following variant of the scan statistic

$$\max_{\mathbb{S} \in \mathbb{S}} \frac{1}{\sqrt{|\mathbb{S}|}} \sum_{v \in \mathbb{S}} (X_v - \mathbb{E}_0(X_v)). \quad (2)$$

(\mathbb{E}_θ denotes the expectation with respect to F_θ , and for a discrete set \mathcal{S} , $|\mathcal{S}|$ denotes its cardinality.) As argued in (Arias-Castro and Grimmett, 2013), this test is in a certain sense asymptotically equivalent to the generalized likelihood ratio test.

1.3 Calibration by permutation

Suppose we are considering a test that rejects the null for large values of a test statistic $T(\mathbf{X})$ where $\mathbf{X} = (X_v, v \in \mathcal{V})$. Let $\mathbf{x} = (x_v, v \in \mathcal{V})$ the observed value of \mathbf{X} . If we were to know the null distribution F_0 , we would return the P-value as $\mathbb{P}_0(T(\mathbf{X}) \geq T(\mathbf{x}))$. In practice, even with the knowledge of F_0 computing the exact P-value might be difficult, but one can approximate it to an arbitrary accuracy and estimate it by Monte Carlo simulation.

Ignoring computational constraints for the moment, calibration by permutation amounts to computing $T(\mathbf{x}_\pi)$ for all $\pi \in \mathcal{V}!$, where $\mathcal{V}!$ denotes the set of all permutations of \mathcal{V} and $\mathbf{x}_\pi = (x_{\pi(v)}, v \in \mathcal{V})$ is the permuted data. We then return the P-value

$$\frac{1}{|\mathcal{V}|!} |\{\pi \in \mathcal{V}! : T(\mathbf{x}_\pi) \geq T(\mathbf{x})\}|$$

and the rejection decision is based on this value. Let $M = |\{T(\mathbf{x}_\pi) : \pi \in \mathcal{V}!\}|$. If there are no multiplicities, meaning $M = \mathcal{V}!$, it can be shown such tests are exact and that under the null the P-value has a (discrete) uniform distribution on $\{1/M, 2/M, \dots, 1\}$. Otherwise the test will be slightly

⁵A family of densities ($f_\theta : \theta \in \Theta$), where $\Theta \subset \mathbb{R}$, has the monotone likelihood ratio property if $f_{\theta'}(x)/f_\theta(x)$ is increasing in x when $\theta' > \theta$.

conservative (Lehmann and Romano, 2005). In practice, the number of permutations is very large (as $|\mathcal{V}|! = |\mathcal{V}|!$) and the P-value is estimated by simulation (by uniform sampling of permutations).

In our setting, T above will be a form of a scan statistic, similar to the one in (2), which maximizes a standardized sum of data entries over a class \mathbb{S} of possible anomalous sets. When calibrating by permutation we are comparing the value $T(\mathbf{x})$ of this statistic on the original data \mathbf{x} with the corresponding value $T(\mathbf{x}_\pi)$ on permuted data \mathbf{x}_π . This is only sensible if the class \mathbb{S} has some structure, and in particular it cannot be invariant under permutations. In this paper we consider what is perhaps the simplest such class, which is the class of intervals

$$\mathcal{V} = \{1, \dots, N\} \text{ and } \mathbb{S} = \{\{a, \dots, b\} : 1 \leq a \leq b \leq N\}.$$

In the next section we elaborate on other possible structural constraints, and the theoretical approach we develop can be used to study the calibration by permutation in those settings as well.

Assuming T has been chosen, we define the oracle scan test as the scan test calibrated with full knowledge of the null distribution by Monte Carlo simulation, and the permutation scan test as the scan test calibrated by permutation as explained above.

Contribution 1: *We characterize the performance of the permutation scan test in the context of the exponential family, concluding that it has as much asymptotic power as the oracle scan test (Theorem 1).*

We note that permutation tests are known to perform this well in classical two-sample testing (Lehmann and Romano, 2005). However, in the context of the scan test, we are only aware of one other paper, that of Walther (2010), that develops theory for the permutation scan test. This is done in the context of binary data (a Bernoulli model). Our analysis extends the theory to any natural exponential model as described in Section 1.2 (which also includes the binary case). This requires a different set of tools.

1.4 Scanning the ranks

As explained earlier, when calibrating by permutation the computation of the scan statistic T must be done for a large enough number of permutations of the original dataset. Even though this is done for only a relatively small number of permutations, that number is often chosen in the hundreds, if not thousands, meaning that the procedure requires the computation of that many scans. Even if the computation (in fact, approximation) of the scan statistic is done in linear time this can be rather time consuming. Furthermore, for a new instantiation of the data the whole procedure must be undertaken anew. The computational burden of doing so may be prohibitive in some practical situations, for instance, when monitoring a sensor network in real-time.

To mitigate those drawbacks we propose instead a rank-based approach, which avoids the expensive calibration by permutation. The procedure amounts to simply replacing the observations with their ranks⁶ before scanning, so that we end up scanning the ranks instead of the original values. If ties in the ranks are broken randomly the resulting test statistic is distribution-free and therefore can be calibrated by Monte Carlo simulation requiring only the knowledge of the data size (which is $N \equiv |\mathcal{V}|$ in our context). In terms of computational complexity this procedure is as complex as the implementation of a scan test when the null distribution is fully known so there is no computational disadvantage in using ranks. In fact faster implementations might be possible by taking advantage of the discrete nature of the ranks and avoiding floating-point algebra, but these algorithmic considerations are beyond the scope of this paper.

⁶Throughout, the observations are ranked in increasing order of magnitude.

Contribution 2: *We establish the performance of the rank scan test (Theorem 2 and Proposition 3). In the context of the exponential family we show that it has nearly as much asymptotic power as the oracle scan test (Proposition 2).*

This result is remarkable in the sense that the scan test can be completely calibrated before any data has been observed, and yet attain essentially the same power as the optimal test with full knowledge of the statistical model. Such a procedure is very natural (albeit distinct) given the classical literature on nonparametric tests (Hettmansperger, 1984), and rank tests such as Wilcoxon’s are known to perform this well in classical two-sample testing (Hettmansperger, 1984; Lehmann and Romano, 2005).

Our results allow us to precisely quantify how much (asymptotic) power is lost when using the rank scan test versus the oracle scan test. For example, in the normal means model the rank scan test requires a signal magnitude 1.023 times larger than the regular scan test to be asymptotically powerful against anomalous sets that are not too small.

1.5 Structured anomalies

Naturally, the intrinsic difficulty of the detection task depends not only on the data distribution, but also on the complexity of the class of anomalous sets \mathbb{S} . Furthermore, for the permutation or rank-based approaches to be sensible this class must have some structure and not be invariant under permutations, as seen above. In several scenarios structural assumptions on such classes arise very naturally. For instance, grid-like networks are an important special case, arising in applications such as signal and image processing (where the signals are typically regularly sampled) and sensor networks deployed for the monitoring of some geographical area, for example. This situation is considered in great generality and from different perspectives in (Arias-Castro et al., 2011, 2005; Cai and Yuan, 2014; Desolneux et al., 2003; Hall and Jin, 2010; Perone Pacifico et al., 2004; Walther, 2010). Also, the distribution of the corresponding scan statistic (2) and variants has been studied in a number of places (Boutsikas and Koutras, 2006; Jiang, 2002; Kabluchko, 2011; Sharpnack and Arias-Castro, 2014; Siegmund and Venkatraman, 1995).

The simplest and most emblematic setting is that of detecting an interval in a one-dimensional regularly sampled signal, that was highlighted above. However, the principles underlying the detection of intervals can be used for the detection of much more general anomaly classes. As shown in (Arias-Castro et al., 2011), similar results apply to a general (nonparametric) class \mathbb{S} of blob-like (‘thick’) sets \mathcal{S} when \mathcal{V} is a grid-like set of arbitrary finite dimension, although the scanning is done over an appropriate approximating net for \mathbb{S} (instead of the entire class \mathbb{S}). Furthermore, these results generalize to one-parameter exponential models, beyond the commonly assumed normal location model, as long as the sets $\mathcal{S} \in \mathbb{S}$ are sufficiently large (poly-logarithmic in N). Other papers that develop theory for different environments include (Addario-Berry et al., 2010; Arias-Castro et al., 2008; Sharpnack and Singh, 2010; Sharpnack et al., 2013; Zhao and Saligrama, 2009). Variants of this detection problem have been suggested, and the applied literature is quite extensive. We refer the reader to (Arias-Castro et al., 2011) and references therein.

Since the main motivation of our work is to develop methods and theory for the scenario when the distributions are unknown/unspecified we focus exclusively on the detection of intervals, for the sake of clarity and simplicity. Nevertheless our techniques and results apply naturally to more general anomaly classes (e.g., rectangles in two or more dimensions, or even blob-like subsets). The key to these generalizations are proper concentration inequalities for sampling without replacement, namely Lemmas 2 and 4, and a geometric characterization of the anomaly class in terms of an approximating net akin to Lemma 1. The latter characterization is heavily dependent on the class of anomalous sets under consideration, as described in the preceding paragraph. Furthermore,

although it is possible to study a version of the test than scans over all possible anomalous sets, we choose to study a scan test restricted to an approximating net because of the following advantages: the analysis is simpler as it does not require the use of chaining to achieve tight constants; it is applicable in more general settings, in particular when the class \mathbb{S} is nonparametric; it is computationally advantageous as it gives rise to fast implementations.

1.6 Content and notation

The rest of the paper is organized as follows. In Section 2 we consider the case when the null distribution is known. This section is expository, introducing the reader to the basic proof techniques that are used, for example, in (Arias-Castro et al., 2011), to establish the performance of the scan statistic when calibrated with full knowledge of the null distribution — the oracle scan test, as we called it here. To keep the exposition simple, and to avoid repeating the substantially more complex arguments detailed in that paper and others, we focus on the problem of detecting an interval in a one-dimensional lattice. This allows us to set the foundation and discover what the performance bounds for the scan test in this case rely on. In Section 3 we consider the same setting and instead calibrate the scan statistic by permutation. In Section 4 we consider the same setting and instead scan the ranks. In both cases, our analysis relies on concentration inequalities for sums of random variables obtained from sampling without replacement from a finite set of reals, already established in the seminal paper of Hoeffding (1963). In Section 5 we perform some simulations to numerically quantify how much is lost in finite samples when calibrating by permutation or when using ranks. We also compare our methodology with the method of Cai et al. (2012), on simulated data, and also on a real dataset from genomics. Section 6 is a brief discussion. Except for the expository derivations in Section 2, the technical arguments are gathered in Section 7.

2 When the null distribution is known

This section is meant to introduce the reader to the techniques underlying the performance bounds developed in (Arias-Castro et al., 2011, 2005) for the scan statistic (and variants) when the null distribution is known. These provide a stepping stone for our results in regards to permutation and rank scan tests. We detail the setting of detecting an interval of unknown length in a one-dimensional lattice. Therefore, as in Section 1.3, consider the setting where

$$\mathcal{V} = \{1, \dots, N\} \text{ and } \mathbb{S} = \{[a, b] : 1 \leq a \leq b \leq N\}.$$

We begin by considering the normal model — $X_v \sim \mathcal{N}(\theta_v, 1)$ are independent — and explain later on how to generalize the arguments to an arbitrary exponential model as described in Section 1.2. We are interested in testing

$$H_0 : \theta_v = 0, \forall v \in \mathcal{V} \quad \text{versus} \quad H_1 : \exists \mathcal{S} \in \mathbb{S} : \frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \sqrt{2 \log(N)/|\mathcal{S}|}, \quad (3)$$

where $\tau > 0$ is fixed. We consider this problem from a minimax perspective. It is shown in (Arias-Castro et al., 2005) that, if $\tau < 1$, then any test with level α has power at most $\beta(\alpha, N)$, with $\beta(\alpha, N) \rightarrow \alpha$ as $N \rightarrow \infty$. In other words, in the large-sample limit, no test can do better than random guessing — the test that rejects with probability α regardless of the data. On the other hand, if $\tau > 1$, then for any level $\alpha > 0$ there exists a test with level α and power $\beta(\alpha, N) \rightarrow 1$ as $N \rightarrow \infty$. In particular, such a test can be constructed using a form of scanning over an approximating net, as explained in the rest of this section.

Step 1: Construction of an approximating net. Instead of scanning over \mathbb{S} we will scan over a subclass of intervals \mathbb{S}_b , where $0 \leq b \leq N$ is an integer to be specified later on. This brings both computational and analytical advantages over scanning all sets in \mathbb{S} as discussed in Section 1.5. Such a subclass must satisfy two important properties, namely have cardinality significantly smaller than \mathbb{S} , and be such that any element $\mathcal{S} \in \mathbb{S}$ can be “well approximated” by an element of $\mathcal{S}^* \in \mathbb{S}_b$. By well approximated we mean that $\rho(\mathcal{S}, \mathcal{S}^*) \approx 1$ where

$$\rho(\mathcal{S}, \mathcal{S}^*) := \frac{|\mathcal{S} \cap \mathcal{S}^*|}{\sqrt{|\mathcal{S}| |\mathcal{S}^*|}},$$

is a measure of similarity of two sets. We use an approximating net similar to that of (Arias-Castro et al., 2005); see (Sharpnack and Arias-Castro, 2014) for an alternative construction.

To simplify the presentation assume N is a power of 2 (namely $N = 2^q$ for some integer q). Let \mathbb{D}_j denote the class of dyadic intervals at scale j , meaning of the form $\mathcal{S} = [1 + k2^j, (k+1)2^j] \subset \mathcal{V}$ with j and k nonnegative integers. Let $\mathbb{D}_{j,0}$ denote the class of intervals of the form $S \cup S'$ with $S, S' \in \mathbb{D}_{j-1}$. Note that $\mathbb{D}_j \subset \mathbb{D}_{j,0}$. Then, for $1 \leq k < b$, let $\mathbb{D}_{j,k}$ be the class of intervals of \mathcal{V} of the form $S_{\text{left}} \cup S \cup S_{\text{right}}$, where $S \in \mathbb{D}_{j,k-1}$ while S_{left} (resp. S_{right}) is adjacent to S on the left (resp. right) and is either empty or in \mathbb{D}_{j-k} . Note that $\mathbb{D}_{j,k-1} \subset \mathbb{D}_{j,k}$ by construction. In the last step, $\mathbb{D}_{j,b}$ is of the same form as before, only the appended intervals S_{left} and S_{right} are either empty, or in \mathbb{D}_{j-b+1} . Finally, define $\mathbb{S}_b = \bigcup_j \mathbb{D}_{j,b}$.

We can prove the following result for this approximating net, using similar arguments to those of Arias-Castro et al. (2005).

Lemma 1. *The subclass $\mathbb{S}_b \subset \mathbb{S}$ has cardinality at most $N4^{b+1}$ and is such that for any element $\mathcal{S} \in \mathbb{S}$ there is an element $\mathcal{S}^* \in \mathbb{S}_b$ satisfying $\mathcal{S} \subset \mathcal{S}^*$ and $\rho(\mathcal{S}, \mathcal{S}^*) \geq (1 + 2^{-b+2})^{-1/2}$.*

Remark 1. It is easy to see that the subclass \mathbb{S}_b can be scanned in $O(Nb4^b)$ operations — this is implicit in (Arias-Castro et al., 2005). Indeed, we start by observing that scanning all dyadic intervals can be done in $O(N)$ operations by recursion, starting from the smallest intervals and moving up (in scale) to larger intervals. We then conclude by realizing that each interval in \mathbb{S}_b is the union of at most $2b+2$ dyadic intervals.

Step 2: Definition of the scan test. We consider a test based on scanning only the intervals in \mathbb{S}_b . This test rejects the null if

$$\max_{\mathcal{S} \in \mathbb{S}_b} Y_{\mathcal{S}} \geq \sqrt{2(1+\eta) \log N} \quad \text{with} \quad Y_{\mathcal{S}} := \frac{1}{\sqrt{|\mathcal{S}|}} \sum_{v \in \mathcal{S}} X_v, \quad (4)$$

where $\eta > 0$ satisfies $\eta \rightarrow 0$ and $\eta \log(N) \rightarrow \infty$. (The reason for these conditions will become clear shortly.)

Step 3: Under the null hypothesis. By the union bound, we have

$$\begin{aligned} \mathbb{P}_0 \left(\max_{\mathcal{S} \in \mathbb{S}_b} Y_{\mathcal{S}} \geq \sqrt{2(1+\eta) \log N} \right) &\leq \sum_{\mathcal{S} \in \mathbb{S}_b} \mathbb{P}_0 \left(Y_{\mathcal{S}} \geq \sqrt{2(1+\eta) \log N} \right) \\ &\leq |\mathbb{S}_b| \bar{\Phi} \left(\sqrt{2(1+\eta) \log N} \right), \end{aligned}$$

where Φ denotes the standard normal distribution function and $\bar{\Phi} = 1 - \Phi$ denotes the corresponding survival function. We have the well-known bound on Mill’s ratio:

$$\bar{\Phi}(x) \leq e^{-x^2/2}, \quad \forall x \geq 0. \quad (5)$$

Therefore we get

$$\mathbb{P}_0 \left(\max_{\mathcal{S} \in \mathbb{S}_b} Y_{\mathcal{S}} \geq \sqrt{2(1+\eta) \log N} \right) \leq N 4^{b+1} N^{-(1+\eta)} = N^{-\eta} 4^{b+1} .$$

We choose $b = \frac{1}{2}\eta \log(N)/\log(4)$. With our assumption that $\eta \log N \rightarrow \infty$, this makes the last expression tend to zero as $N \rightarrow \infty$. (It also implies that $b \rightarrow \infty$, which we use later on.) We conclude the test in (4) has level tending to 0 as $N \rightarrow \infty$.

Step 4: Under the alternative. We now show that the power of this test tends to 1 when $\tau > 1$. Let \mathcal{S} denote the anomalous interval. Referring to Lemma 1, there is a set $\mathcal{S}^* \in \mathbb{S}_b$ such that $\rho(\mathcal{S}, \mathcal{S}^*) \geq (1 + 2^{-b+2})^{-1/2}$, so that $\rho(\mathcal{S}, \mathcal{S}^*) = 1 + o(1)$ since $b \rightarrow \infty$. Furthermore $Y_{\mathcal{S}^*}$ is normal with mean at least $\rho(\mathcal{S}, \mathcal{S}^*)\tau\sqrt{2\log N}$ and variance 1. We thus have

$$\mathbb{P} \left(Y_{\mathcal{S}^*} \geq \sqrt{2(1+\eta) \log N} \right) \geq \bar{\Phi}(\xi) ,$$

where

$$\begin{aligned} \xi &:= \sqrt{2(1+\eta) \log N} - \rho(\mathcal{S}, \mathcal{S}^*)\tau\sqrt{2\log N} \\ &= \sqrt{2(1+\eta) \log N} (1 - (1 + o(1))\tau/\sqrt{1+\eta}) \\ &\sim -(\tau - 1)\sqrt{2\log N} \rightarrow -\infty , \end{aligned}$$

where we used the fact that $\tau > 1$ is fixed and $\eta \rightarrow 0$. We conclude that the test in (4) has power tending to 1 as $N \rightarrow \infty$. In conclusion, we have shown the following result.

Proposition 1 (Arias-Castro et al. (2005)). *Refer to the hypothesis testing problem in (3). The test defined in (4), with $\eta = \eta_N \rightarrow 0$, $\eta_N \log N \rightarrow \infty$ and $b = b_N = \frac{1}{2}\eta_N \log N$, has level converging of 0 as $N \rightarrow \infty$. Moreover when $\tau > 1$ it has power converging to 1 as $N \rightarrow \infty$.*

We remark that, in principle, we may choose any $b = b_N \rightarrow \infty$ such that $b_N/\log N \rightarrow 0$. From Remark 1 the computational complexity of the resulting scan test is of order $O(Nb_N 4^{b_N})$. For example, $b_N \sim \log \log N$ is a valid choice and the resulting scan test runs in $O(N \text{polylog}(N))$ time.

2.1 Generalizations

The arguments just given for the setting of detecting an anomalous interval under a normal location model can be generalized to the problem of detecting other classes of subsets under other kinds of distributional models. We briefly explain how this is done. (Note that these generalizations can be combined.)

Other classes of anomalous subsets For a given detection problem, specified by a set of nodes \mathcal{V} and a class of subsets $\mathbb{S} \subset 2^{\mathcal{V}}$, the arguments above continue to apply if one is able to construct an appropriate approximating net as in Lemma 1. This is done, for example, in (Arias-Castro et al., 2011, 2005) for a wide range of settings. We note that the construction of a net is purely geometrical and/or combinatorial.

Other exponential models To extend the result to an arbitrary (one-parameter, natural) exponential model, we require the equivalent of the tail-bound (5). While such a bound may not apply to a particular exponential model, it does apply asymptotically to large sums of IID variables from that model by Chernoff's bound and a Taylor development of the rate function.

Indeed, recalling the notation introduced in Section 1.2, let $\psi_0(t) = \sup_{\lambda \in [0, \theta^*)} (\lambda t - \log \varphi_0(\lambda))$, which is the rate function of F_0 . By Chernoff's bound, we have

$$\mathbb{P}_0(Y_{\mathcal{S}} \geq y) \leq \exp\left(-|\mathcal{S}|\psi_0(y/|\mathcal{S}|^{-1/2})\right). \quad (6)$$

Assuming without loss of generality that F_0 has zero mean and unit variance, we have

$$\psi_0(t) \geq \frac{1}{2}t^2 + O(t^3), \quad t \rightarrow 0. \quad (7)$$

To see this, note that $\varphi_0(\lambda)$ is infinitely many times differentiable when $\lambda \in [0, \theta^*)$, with $\varphi_0'(0) = \mathbb{E}_0(X) = 0$ and $\varphi_0''(0) = \mathbb{E}_0(X^2) = 1$. Therefore $\varphi_0(\lambda) = 1 + \frac{1}{2}\lambda^2 + O(\lambda^3)$ as $\lambda \rightarrow 0$. For $t \in [0, \theta^*)$, we then have

$$\begin{aligned} \psi_0(t) &= \sup_{\lambda \in [0, \theta^*)} [\lambda t - \log \varphi_0(\lambda)] \geq t^2 - \log \varphi_0(t) \\ &= t^2 - \log\left(1 + \frac{1}{2}t^2 + O(t^3)\right) \geq \frac{1}{2}t^2 + O(t^3), \end{aligned}$$

where we use $\log(1+x) \leq x$. From this we see that our derivations for the normal model apply essentially verbatim if, for some constant $c > 0$, $|\mathcal{S}| \geq c(\log N)^3$ for all $\mathcal{S} \in \mathbb{S}$. Furthermore, it can be seen that the test in (4) is essentially optimal for exponential models, as its performance matches the lower bounds in (Arias-Castro et al., 2011).

3 Calibration by permutation

Having described in detail how a performance bound is established for the scan test variant (4) for the problem of detecting an interval of unknown length, and its extensions to other detection problems, we now clearly see that the key to adapting this analysis to a calibration by permutation is a concentration of measure bound to replace (5) and (6). Since this is the same in any detection setting, we consider as in Section 2 the problem of detecting an interval of unknown length. This time, we impose a minimum and maximum length on the intervals

$$\mathbb{S} = \left\{ \{a, \dots, b\} : 1 \leq a < b \leq N, 2^{q_l} \leq b - a \leq 2^{q_u} \right\}. \quad (8)$$

Indeed, when calibrating the scan test by permutation, we necessarily have to assume nontrivial upper and lower bounds on the size of an anomalous interval. To see this consider intervals of length one. Then the value of the scan for any permutation of the data is the same. By symmetry the same reasoning applies for intervals of length $N - 1$.

We consider essentially the same form of the scan statistic (2) as before, but replace $\mathbb{E}_0(X_v)$ (which we do not have access to) by $\bar{X} = \frac{1}{N} \sum_{v \in \mathcal{V}} X_v$ and scan over an approximating net. We restrict the approximating net to match the class of intervals defined in (8) (but still call it \mathbb{S}_b for simplicity). Specifically we only keep an element $\mathcal{S}^* \in \mathbb{S}_b$ if there is $S \in \mathbb{S}$ such that $\rho(\mathcal{S}, \mathcal{S}^*) \geq (1 + 2^{-b+2})^{-1/2}$. This ensures that the statements in Lemma 1 still hold, and also that $|\mathcal{S}^*| \geq 2^{q_l}/(1 + 2^{-b+2})$ for all $\mathcal{S}^* \in \mathbb{S}_b$. In detail, with $\mathbf{x} = (x_v, v \in \mathcal{V})$ denoting the observed data, we define

$$\text{SCAN}(\mathbf{x}) = \max_{\mathcal{S} \in \mathbb{S}_b} \left(Y_{\mathcal{S}}(\mathbf{x}) - \sqrt{|\mathcal{S}|} \bar{x} \right), \quad Y_{\mathcal{S}}(\mathbf{x}) := \frac{1}{\sqrt{|\mathcal{S}|}} \sum_{v \in \mathcal{S}} x_v, \quad (9)$$

The test rejects the null at significance level $\alpha \in (0, 1)$ when

$$\mathfrak{P}(\mathbf{x}) := \frac{1}{|\mathcal{V}|!} \left| \left\{ \pi \in \mathcal{V}! : \text{SCAN}(\mathbf{x}_{\pi}) \geq \text{SCAN}(\mathbf{x}) \right\} \right| \leq \alpha, \quad (10)$$

where $\mathfrak{P}(\mathbf{x})$ is the permutation P-value.

Theorem 1. *Refer to the hypothesis testing problem in (3) and assume F_0 has zero mean and variance one. Consider the test that rejects the null if $\mathfrak{P}(\mathbf{X}) \leq \alpha$, where \mathfrak{P} is defined in (10), with $b = b_N \rightarrow \infty$ and $b_N/\log N \rightarrow 0$ at $n \rightarrow \infty$. This test has level at most α . Furthermore, assume that under the alternative the anomalous set \mathcal{S} belongs to \mathbb{S} defined in (8) with $q_l - 3\log_2 \log N \rightarrow +\infty$ and $q_u - \log_2 N \rightarrow -\infty$ as $N \rightarrow \infty$. This test has power converging to 1 as $N \rightarrow \infty$ when*

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \sqrt{2\log(N)/|\mathcal{S}|}, \quad \text{with } \tau > 1 \text{ fixed,}$$

provided that either F_0 has compact support or $\max_v \theta_v \leq \tilde{\theta} < \theta_$ for some fixed $\tilde{\theta} > 0$.*

The headline here is that a calibration by permutation has as much asymptotic power as a calibration by Monte Carlo with full knowledge of the null distribution (to first-order accuracy). This is (qualitatively) in line with what is known in classical settings (Lehmann and Romano, 2005). Note that this testing procedure makes no assumptions about F_0 or about the existence of an underlying exponential model.

Remark 2. The assumption that F_0 has zero mean and variance one is without any loss of generality, and merely for clarity of presentation. In general, the permutation-based test is asymptotically powerful under the alternative if there is a set $\mathcal{S} \in \mathbb{S}$ such that

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \frac{1}{\sigma_0} \sqrt{2\log(N)/|\mathcal{S}|}, \quad \text{with } \tau > 1 \text{ fixed,}$$

where σ_0^2 denotes the variance of F_0 .

The conditions required here allow \mathbb{S} to be any class of intervals of lengths between $(\log N)^{3+a}$ and $o(N)$, for any $a > 0$ fixed. This includes the most interesting cases of intervals not too short and also not too long. In fact, for certain families of distributions removing from consideration very small intervals is essential and cannot be avoided.

Example 1. For instance consider the Bernoulli model, where $X_v \sim \text{Bernoulli}(1/2)$, for all $v \in \mathcal{V}$ under the null, and $X_v \sim \text{Bernoulli}(1)$, for all $v \in \mathcal{S}$ when \mathcal{S} is anomalous. Even under the null we will encounter a run of ones of length $\sim \log_2 N$ (the famous Erdős–Rényi Law) with positive probability. Therefore in this case the scan test, calibrated by Monte Carlo or permutation, is powerless for detection of intervals of length $\frac{1}{2} \log_2 N$. In fact, it can be shown that no test has any power in that case.

Note that, when calibrating a test by permutation there are essentially two sources of randomness. The randomness intrinsic to the data \mathbf{X} , and the randomness induced by the permutation. In particular, if we regard π as a uniform random variable over the set of possible permutations $\mathcal{V}!$ the P-value of the test can be re-written as $\mathfrak{P}(\mathbf{X}) = \mathbb{P}(\text{SCAN}(\mathbf{X}_\pi) \geq \text{SCAN}(\mathbf{X}))$. Under the null hypothesis the argument is classic: for any given permutation π , the distribution of \mathbf{X} is identical to the distribution of \mathbf{X}_π , therefore $\text{SCAN}(\mathbf{X})$ is conditionally uniformly distributed in $\{\text{SCAN}(\mathbf{X}_\pi) : \pi \in \mathcal{V}!\}$ (with multiplicities). The bulk of the effort in the proof is to characterize the behavior of the test under the alternative. The first step is to, conditionally on the data \mathbf{X} , “remove” the randomness in π . Realizing that for any \mathcal{S} , $\sum_{v \in \mathcal{S}} X_\pi(v)$ is simply a sum of elements sampled without replacement from \mathbf{X} , we are able to use a concentration inequality for sampling without replacement to upper-bound the P-value by an expression involving $\text{SCAN}(\mathbf{X})$, the sample mean and variance of \mathbf{X} , and $\max_v X_v$. The remainder of the proof consists in controlling those terms for the exponential model.

For technical reasons, we place an upper bound $\tilde{\theta}$ on the nonzero θ_v ’s to streamline the proof arguments and be able to control $\max_v X_v$. However, note that this condition is not a simple

artifact of the proof technique and its removal will invalidate the statement. A way around this assumption is to state the result in terms of $\min_{v \in \mathcal{S}} \theta_v$ instead of $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v$ and use censoring prior to scanning (see the discussion in Section 6).

4 Scanning the ranks

Having observed $\mathbf{x} = (x_v, v \in \mathcal{V})$, scanning the ranks amounts to replacing every observation with its rank among all the observations, and computing the scan (9). We call this the *rank scan*. As for all rank-based methods, the null distribution is the permutation distribution when there are no ties.

- When there are no ties with probability one, calibration of the distribution of the test statistic is determined by the data size N , and therefore the test can be calibrated by Monte Carlo simulation before data is observed.
- When there are ties the rank scan test can be also calibrated by permutation. If one breaks ties using the average rank then calibration must be done anew for any given dataset. A much better alternative is to break ties randomly so that we are back in the first case, and can calibrate the test before seeing the data. The latter option is computationally superior and is the one we analyze.

In summary, the rank scan test is computationally more advantageous, when compared with the test of the previous section, calibrated by permutation. An additional advantage of the rank scan is its robustness to outliers — although the permutation scan after censoring (discussed in Section 6) is also robust to outliers. See Section 5 for implementation issues and a computational complexity analysis.

Formally, let $\mathbf{x} = (x_v, v \in \mathcal{V})$ denote the observations as before, and for every $v \in \mathcal{V}$, let r_v be the rank (in increasing order) of x_v in \mathbf{x} , where ties are broken randomly, and let $\mathbf{r} = (r_v, v \in \mathcal{V})$ be the vector of ranks. The rank scan test returns the P-value $\mathfrak{P}(\mathbf{r})$ defined in (10).

Because the rank scan test is naturally regarded as a kind of permutation scan test, we assume similarly upper and lower bounds on the size of the anomalous set as in Section 3. The first result we present is rather general, and it is not particular to the exponential family and applies to the general setting in Section 2.1. For rank-based procedures the performance will depend naturally on the ability to rank correctly an anomalous observation against a normal one. This is naturally captured by the following quantity:

$$p_v = \mathbb{P}(Y > X) + \frac{1}{2} \mathbb{P}(Y = X), \quad \text{where } X \sim F_0 \text{ and } Y \sim F_v \text{ are independent.} \quad (11)$$

The larger p_v is the higher is the probability of ranking the two observations correctly.

Theorem 2. *Refer to the hypothesis testing problem in Section 2.1 and consider the test that rejects the null if $\mathfrak{P}(\mathbf{R}) \leq \alpha$, where \mathfrak{P} is defined in (10), with $b = b_N \rightarrow \infty$ and $b_N / \log N \rightarrow 0$. This test has level at most α . Furthermore this test has power converging to 1 as $N \rightarrow \infty$ provided*

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v \geq \frac{1}{2} + \tau \sqrt{2 \log(N) / |\mathcal{S}|}, \quad \text{with } \tau > \frac{1}{2\sqrt{3}} \text{ fixed,}$$

and \mathcal{S} belongs to \mathbb{S} defined in (8) with $q_l - \log_2 \log N \rightarrow +\infty$ and $q_u - \log_2 N \rightarrow -\infty$ as $N \rightarrow \infty$.

This result characterizes the performance of the rank scan test for general distributions (actually we do not even need to assume F_v stochastically dominates F_0). To get a better sense of this result and be able to compare it with the previous theorem it is useful to consider the particular case of the exponential model. Define

$$\Upsilon_0 = \frac{1}{2} \mathbb{E}[\max(X, Y)] , \quad (12)$$

where $X, Y \sim F_0$ and independent.

Proposition 2. *Refer to the hypothesis testing problem in (3), assume F_0 has zero mean and variance one, and refer to the test in Theorem 2. The test has level at most α . Moreover, it has power converging to 1 as $N \rightarrow \infty$ when*

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \sqrt{2 \log(N)/|\mathcal{S}|} , \quad \text{with } \tau > \frac{1}{2\sqrt{3}\Upsilon_0} \text{ fixed.}$$

The headline here is that the rank scan requires a signal amplitude which is $1/(2\sqrt{3}\Upsilon_0)$ larger than what is required of the regular scan test calibrated by Monte Carlo with full knowledge of the null distribution. This is (qualitatively) in line with similar results in more classical settings (Hettmansperger, 1984). For the normal location model, we find that $1/(2\sqrt{3}\Upsilon_0) = \sqrt{\pi/3} \approx 1.023$, so the detection threshold of rank scan is almost the same as that of the regular scan test — see the Appendix 7.5.2 for details. Note that $\Upsilon_0 \leq 1/(2\sqrt{3})$ (otherwise this would contradict the known minimax lower bounds) and that equality is attained if and only if F_0 is the uniform distribution.⁷

Remark 3. As in the case of Theorem 2 the assumption on the moments of F_0 are used only for clarity of presentation. In general, the permutation-based test is asymptotically powerful under the alternative if there is a set $\mathcal{S} \in \mathbb{S}$ such that

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \sqrt{2 \log(N)/|\mathcal{S}|} , \quad \text{with } \tau > \frac{\tau}{2\sqrt{3}(\Upsilon_0 - \mu_0/2)} \text{ fixed,}$$

where μ_0 denotes the mean of F_0 .

The proof of Theorem 2 starts essentially as that of Theorem 1. Under the alternative the P-value is upper bounded by an expression involving $\text{SCAN}(\mathbf{R})$. Control of this term is more complicated than that of $\text{SCAN}(\mathbf{X})$ in the previous theorem, since the elements of \mathbf{R} are not independent, but can be done by controlling the first two moments of \mathbf{R} . For Proposition 2 we note that for the exponential model one can relate $p_v \equiv p_{\theta_v}$ to θ_v by a Taylor expansion around zero, concluding the proof.

Small and very small intervals

The conditions of Theorem 2 allow for dealing with intervals of length of order (strictly) larger than $\log N$. We give here results that encompass the scenario where the interval might be of smaller length. To keep the discussion simple we consider the class of intervals of a fixed size $|\mathcal{S}| = k$ under the alternative, and explain later how this result is generalized for a class of intervals of different sizes. In this situation there is no need to consider an approximating net and we simply scan over the entire class, denoted by \mathbb{S} . Recall the definition of the permutation P-value (10).

Proposition 3. *Refer to the hypothesis testing problem in Section 2.1 and consider the test that rejects the null if $\mathfrak{P}(\mathbf{R}) \leq \alpha$. Then the test has level at most α and power converging to 1 as $N \rightarrow \infty$ provided there is an interval \mathcal{S} of length k such that*

⁷This is based on a personal communication from Richard J. Samworth and Tengyao Wang, who got interested in this question after one of the present authors presented this work at Cambridge University.

- (i) $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v = 1 - o(N^{-2/k})$ when $2 < k = o(\log N)$; or
- (ii) $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v > 1 - \frac{1}{2} \exp(-\frac{c+1}{c})$ when $k = c \log N$ for some $c > 0$ fixed.

Theorem 2 and Proposition 3 together cover essentially all interval sizes which are $o(N)$. Theorem 2 covers the case of larger intervals, in which case $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v$ can go to $1/2$ provided it does not converge too fast, and the test is still powerful asymptotically. In Proposition 3, a sufficient condition for an asymptotically powerful test is that $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v$ goes to 1 at a certain rate when the size of the anomalous interval is $o(\log N)$. If the interval size is $c \log N$ with $c > 0$ arbitrary the rank test is asymptotically powerful when $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v$ is greater than a constant (strictly larger than $1/2$) depending on c .

Extending this result to the exponential model is not possible without additional knowledge of the family of distributions, as having p_{θ_v} bounded away from $1/2$ implies θ_v is bounded away from 0. As an example, consider the normal means model when $k = o(\log N)$. In this case, we have

$$p_\theta = \Phi(-\theta/\sqrt{2}) \geq 1 - \frac{1}{2} e^{-\theta^2/4}.$$

Hence, whenever $\frac{1}{2} e^{-\theta^2/4} = o(N^{-2/k})$, the condition in the proposition is met. This is satisfied when

$$\theta = \tau \sqrt{2 \log(N)/k}, \quad \text{with } \tau > 2 \text{ fixed.} \quad (13)$$

This means that in this case the rank scan requires an amplitude at most two times larger than the regular scan test calibrated with full knowledge of the null distribution.

Finally note that the condition $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v \rightarrow 1$ or $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v > 1 - \frac{1}{2} \exp(-\frac{c+1}{c})$ might not be possible to meet for certain distributions of the exponential family. For instance, in Example 1, $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v = 3/4$, a case not covered by Proposition 3 when the interval size is smaller than $c \log N$ and c is small enough. But this is expected since no test has any power if c is sufficiently small.

Remark 4. Proposition 3 considered the case when the size of the anomalous interval is known. However, we could consider the class of intervals of length greater than 2 and at most \tilde{k} for some given $\tilde{k} = O(\log N)$. In this case we would simply scan the ranks for every fixed interval size up to \tilde{k} and apply a Bonferroni correction to the P-values. Following through the steps of the proof, one can see that the rank scan test would be asymptotically powerful when

- (i') $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v = 1 - o(N \log N)^{-2/|\mathcal{S}|}$ when $2 < |\mathcal{S}| = o(\log N)$; or
- (ii') $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v > 1 - \frac{1}{2} \exp(-\frac{c+1}{c})$ when $|\mathcal{S}| = c \log N$ for some $c > 0$ fixed.

For the normal location model and considering $\tilde{k} = o(\log N)$, we can see that this is satisfied when (13) holds.

5 Numerical experiments

5.1 Computational complexity

We already cited some works where fast (typically approximate) algorithms for scanning various classes of subsets are proposed (Arias-Castro et al., 2005; Neill, 2012; Neill and Moore, 2004; Walther, 2010). For example, as we saw in Lemma 1, Arias-Castro et al. (2005) design an approximating net \mathbb{S}_b for the class of all intervals \mathbb{S} that can be scanned in $O(Nb4^b)$. Furthermore, we saw in Proposition 1 that this procedure achieves the optimal asymptotic power as long as $b = b_N \rightarrow \infty$. For example, if $b_N \asymp \log \log N$, then the computational complexity is of order $(N \text{polylog}(N))$.

In any case, suppose that a scanning algorithm has been chosen and let \mathcal{C}_N denote its computational complexity. The oracle scan test and the rank scan test are then comparable, in that they estimate the null distribution of their respective test statistic by simulation, and this is done only once for each data size N . With this preprocessing already done, the computational complexity of these two procedures is \mathcal{C}_N , the cost of a single scan when applied to data of size N . In contrast, the permutation scan test is much more demanding, in that it requires scanning each of the permuted datasets, and this is done every time the test is applied. Assuming B permutations are sampled at random for calibration purposes, the computational complexity is $B\mathcal{C}_N$, that is, B times that of the oracle or rank variants (not accounting for preprocessing). B is typically chosen in the hundreds ($B = 200$ in our experiments), if not thousands, so the computational burden can be much higher for the permutation test.

5.2 Simulations

We present the results of some basic numerical experiments that we performed to corroborate our theoretical findings in finite samples. We generated the data from the normal location model — where $F_\theta = \mathcal{N}(\theta, 1)$ — which is arguably the most emblematic one-parameter exponential family and a popular model in signal and image processing. We used the regular scan test, calibrated with full knowledge of the null distribution, as a benchmark. The permutation scan test and rank scan test were calibrated by permutation.

The test statistic that we use in our experiments is the scan over all intervals of dyadic length. This subclass of intervals is morally similar to \mathbb{S}_0 (corresponding to $b = 0$) but somewhat richer. This choice allows us to both streamline the implementation and make the computations very fast via one application of the Fast Fourier Transform per dyadic length. In detail, letting \mathbb{S} denote the class of all discrete intervals in \mathcal{V} , this amounts to taking as approximating set

$$\mathbb{S}_{\text{dyad}} = \left\{ \mathcal{S} \in \mathbb{S} : |\mathcal{S}| = 2^j \text{ for some } j \in \mathbb{N} \right\}.$$

As explained earlier, the calibration by permutation and the rank-based approach are valid no matter what subclass of intervals is chosen, and in fact, the same mathematical results apply as long as the subclass is an appropriate approximating net. We encourage the reader to experiment with his/her favorite scanning implementation.

It is easy to see that, for each $\mathcal{S} \in \mathbb{S}$, there is $\mathcal{S}^* \in \mathbb{S}_{\text{dyad}}$ with $\mathcal{S}^* \subset \mathcal{S}$ and $|\mathcal{S}^*| > |\mathcal{S}|/2$. Hence,

$$\min_{\mathcal{S} \in \mathbb{S}} \max_{\mathcal{S}^* \in \mathbb{S}_{\text{dyad}}} \rho(\mathcal{S}, \mathcal{S}^*) \geq 1/\sqrt{2}.$$

A priori, this implies that scanning over \mathbb{S}_{dyad} requires an amplitude $\sqrt{2}$ larger to achieve the same (asymptotic) performance as scanning over \mathbb{S} or a finer approximating set as considered previously. To simplify things, however, in our simulations we took an anomalous interval of dyadic length, so that the detection threshold is in fact the same as before.

We set $N = 2^{15}$ and tried two different lengths for the anomalous interval $|\mathcal{S}| \in \{2^7, 2^{10}\}$. All the nonzero θ_v 's were taken to be equal to

$$\theta_{\mathcal{S}} = t\sqrt{2\log(N)/|\mathcal{S}|} \quad (14)$$

with t varying. The critical values and power are based on 1000 repeats in each case. A level of significance of 0.05 was used. Also, 200 permutations were used for the permutation scan test. The results are presented in Figure 5.2. At least in these small numerical experiments, the three tests behave comparably, with the rank scan slightly dominating the others. Although the last finding is

somewhat surprising, this is a finite-sample effect and is localized in the intermediate power range (around a power of 0.5) and so does not contradict the theory developed earlier. In fact, the three tests achieve power 1 at roughly the same signal amplitude, confirming the theory.

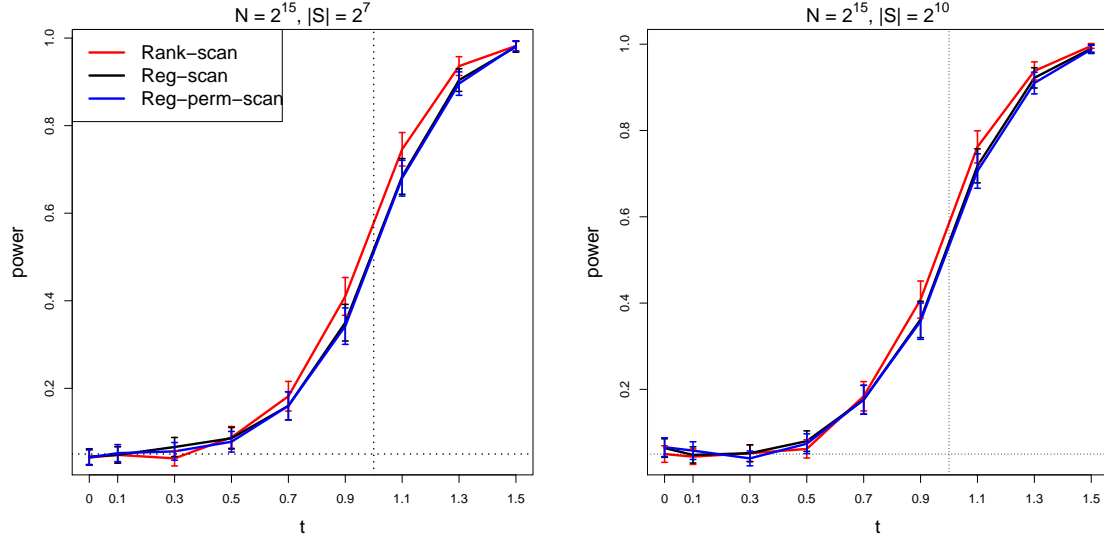


Figure 1: Power curves (with 95% margin of error) for the three tests (all set at level 0.05) as a function of the parameter t in (14): the scan test calibrated with knowledge of the null distribution (black); the permutation scan test (blue); and the rank scan test (red). On the left are the results for $|\mathcal{S}| = 2^7$ and on the right for $|\mathcal{S}| = 2^{10}$. $N = 2^{15}$ in both cases. Each situation was repeated 1,000 times and each time 200 permutations were drawn for calibration. The vertical black dashed line is the minimax boundary for t . The horizontal black dashed line is the significance level 0.05.

5.3 Comparison with RSI

Next, we compare our rank scan with the robust segment identifier (RSI) of [Cai et al. \(2012\)](#). This is a recent method based taking the median over bins of a certain size (a tuning parameter of the method) and then scanning over intervals. Because the median is asymptotically normal, it allows for a calibration that only requires the value of the null density at 0. In turn, one can try to estimate this parameter. Although the method is not distribution-free proper, it appears to be the main contender in the literature. We first compare the two methods on simulated data, for in the context of detection (the problem we considered so far) and in the context of identification (a problem considered in that paper).

Detection In the problem of detection, we compare the performance of the rank scan test and RSI with bin size $m \in \{10, 20\}$ in normal data. To turn RSI into a test, we reject if it detects any anomalous interval. In the simulation, we set sample size $N = 50,000$ and considered the case where there is only one signal interval with known length $|\mathcal{S}| \in \{100, 1000\}$. The amplitude satisfy (14) as before. We report the empirical power curves (based on 100 repeats) in Figure 2.

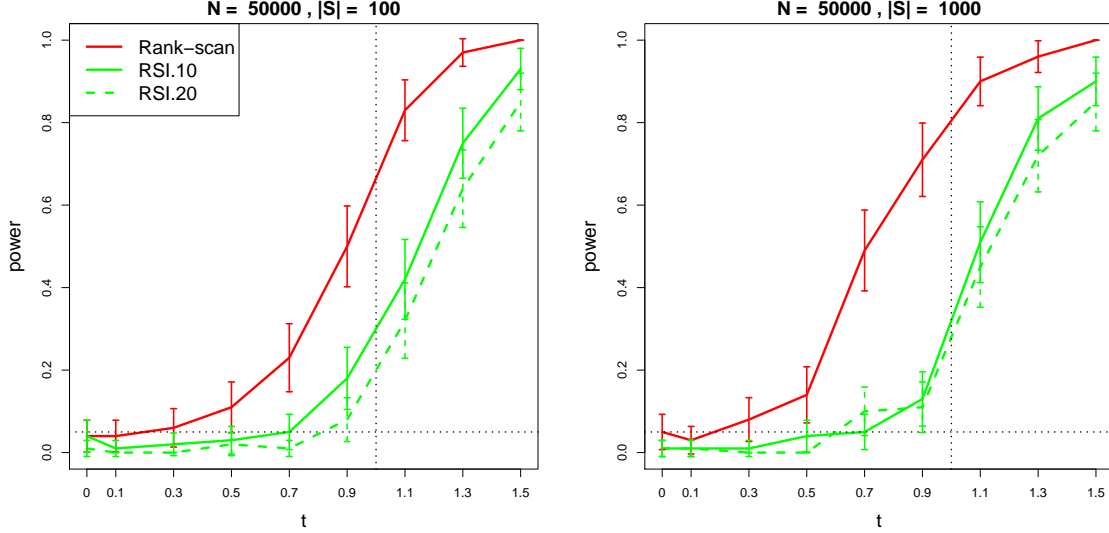


Figure 2: Power curves (with 95% margin of error) for the three tests as a function of the parameter t in (14): the rank scan test (red); RSI with bin size 10 (solid green); and RSI with bin size 20 (dashed green). The rank scan test is set at level 0.05 and its critical value is from 1000 repeats. On the left are the results for $|\mathcal{S}| = 100$ and on the right for $|\mathcal{S}| = 1000$. $N = 50,000$ in both cases. Each situation was repeated 100 times. The vertical black dashed line is the minimax threshold for t . The horizontal black dashed line is the significance level 0.05.

To be fair, both methods only scan candidate signal intervals of length $|\mathcal{S}|$. The rank scan is calibrated as before. For RSI, we set the threshold to $\sqrt{2\log N}$ for the normalized data after localization to better control the family-wise type I error as explained in (Cai et al., 2012). From Figure 2, we can see that RSI is a bit more conservative. In fact, a drawback of RSI is the difficulty to calibrate it correctly.⁸ In any case, the rank scan test outperforms RSI in these simulations.

Identification In the problem of identification, we compare the rank scan and RSI. Although we focused on the problem of detection so far, a scan can be as easily used for testing as for estimation (i.e., identification). Indeed, one sets an identification threshold and extract all the intervals that exceed that threshold. Some post-processing — such as merging significant intervals that intersect or keeping the most significant among significant intervals that intersect — is often applied.

Here, in an effort to be fair, we simply took the procedure of (Cai et al., 2012) — which is essentially the procedure of (Jeng et al., 2010) — but calibrating as we did for testing. Note that this implies a very stringent false identification rate (at the 0.05 testing level this means that the chances that one or more intervals are identified by mistake is 0.05). We then compare its performance to that of the rank scan testing procedure calibrated in the same fashion.

Following (Cai et al., 2012), in the simulation, we set the sample size to $N = 10^4$. We consider a range of null distributions: the standard normal distribution, the t -distribution with 15 degrees of freedom and that with one degree of freedom. In each case, we set the signal mean to $\theta_{\mathcal{S}} \in \{1, 1.5, 2\}$. There are three signal intervals, $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$, starting at positions 1000, 2000, 3000, and having lengths $2^4, 2^5, 2^6$, respectively. We set the threshold for the rank scan test by simulation at a significance level of 0.05. For RSI, we tried several bin sizes, $m \in \{2^3, 2^5\}$. To simplify the computation, both

⁸Of course, it could be calibrated by permutation, but this would make the procedure much more like the permutation scan test (with the same high-computational burden), somewhat far from the intentions of (Cai et al., 2012).

methods only scan dyadic intervals of length at most 2^6 . As in (Cai et al., 2012), we compare their performance in terms of the following dissimilarities

$$D_j = \min_{\hat{\mathcal{S}} \in \hat{\mathbb{S}}} \{1 - \rho(S_j, \hat{\mathcal{S}})\},$$

and the number of false positives, namely

$$O = \{\hat{\mathcal{S}} \in \hat{\mathbb{S}} : \hat{\mathcal{S}} \cap \mathcal{S} = \emptyset, \forall \mathcal{S} \in \mathbb{S}\},$$

where $\hat{\mathbb{S}}$ are the estimated signal intervals.

We report the average and standard deviation (in the parenthesis in the tables below) based on 200 repeats in Tables 1, 2, and 3. We can see that the rank scan method performs better than RSI in when the null distribution is normal and $t(15)$, and it performs similarly to RSI with bin size $m = 2^3$ in $t(1)$. However, when the bin size of RSI is not properly chosen, RSI can perform poorly.

Table 1: Dissimilarity and number of over-selected intervals in $\mathcal{N}(0, 1)$

$\theta_{\mathcal{S}}$	Method	$D_1(\mathcal{S}_1 = 2^4)$	$D_2(\mathcal{S}_2 = 2^5)$	$D_3(\mathcal{S}_3 = 2^6)$	$\#O$
1	Rank Scan	0.734 (0.421)	0.148 (0.284)	0.031 (0.049)	0.000 (0.000)
	RSI($m = 2^3$)	0.916 (0.235)	0.420 (0.406)	0.095 (0.091)	0.065 (0.267)
	RSI($m = 2^5$)	0.998 (0.029)	0.959 (0.144)	0.326 (0.278)	0.130 (0.337)
1.5	Rank Scan	0.167 (0.326)	0.019 (0.044)	0.008 (0.012)	0.000 (0.000)
	RSI($m = 2^3$)	0.593 (0.391)	0.132 (0.033)	0.069 (0.029)	0.080 (0.272)
	RSI($m = 2^5$)	0.980 (0.087)	0.729 (0.284)	0.204 (0.044)	0.025 (0.157)
2	Rank Scan	0.018 (0.051)	0.006 (0.024)	0.004 (0.008)	0.000 (0.000)
	RSI($m = 2^3$)	0.277 (0.226)	0.128 (0.021)	0.064 (0.013)	0.065 (0.247)
	RSI($m = 2^5$)	0.960 (0.122)	0.476 (0.162)	0.193 (0.032)	0.010 (0.100)

5.4 Application to the real data

In this section, we apply the methods to the problem of detecting the copy number variant (CNV) in the context of next generation sequencing data. We compare the rank scan method and RSI on the task of identifying short reads on chromosome 19 of a HapMap Yoruban female sample (NA19240) from the 1000 genomes project (<http://www.1000genomes.org>), which is the same data set used in (Cai et al., 2012). Following standard protocols (Ernst et al., 2011), we extend all the reads to 100 base pairs (BPs). We take 10^6 reads from the whole data set for comparison purposes resulting in 1,281,502 genomic locations.

We tune RSI as done in (Cai et al., 2012), setting the bin size to $m = 400$ and the maximum BPs in a possible CNV to $L = 2^{16}$. Note that (Cai et al., 2012) took $L = 60,000$, which is a bit smaller than 2^{16} . (We chose the latter because we only scan intervals of dyadic length.) To save computational time, in the implementation of the rank scan we group read depths in every 200 positions and take the summation of the read depths for each bin and use that as input (meaning, we rank the sums and scan the ranks). We get the critical value for the rank scan method under the significance level 0.05 from 1000 repeats. In the experiment, we let RSI and the rank scan method only scan dyadic intervals of lengths from 2^1 to 2^{16} .

Table 2: Dissimilarity and number of over-selected intervals in $t(15)$

θ_S	Method	$D_1(\mathcal{S}_1 = 2^4)$	$D_2(\mathcal{S}_2 = 2^5)$	$D_3(\mathcal{S}_3 = 2^6)$	$\#O$
1	Rank Scan	0.806 (0.369)	0.223 (0.354)	0.029 (0.048)	0.000 (0.000)
	RSI($m = 2^3$)	0.926 (0.223)	0.436 (0.406)	0.106 (0.099)	0.050 (0.218)
	RSI($m = 2^5$)	0.996 (0.041)	0.944 (0.168)	0.336 (0.278)	0.125 (0.332)
1.5	Rank Scan	0.232 (0.378)	0.026 (0.079)	0.010 (0.017)	0.000 (0.000)
	RSI($m = 2^3$)	0.554 (0.391)	0.143 (0.112)	0.069 (0.031)	0.075 (0.282)
	RSI($m = 2^5$)	0.992 (0.057)	0.732 (0.286)	0.199 (0.042)	0.020 (0.140)
2	Rank Scan	0.034 (0.097)	0.009 (0.019)	0.005 (0.014)	0.000 (0.000)
	RSI($m = 2^3$)	0.277 (0.220)	0.128 (0.022)	0.063 (0.013)	0.060 (0.238)
	RSI($m = 2^5$)	0.968 (0.107)	0.521 (0.214)	0.192 (0.030)	0.010 (0.100)

Table 3: Dissimilarity and number of over-selected intervals in $t(1)$

θ_S	Method	$D_1(\mathcal{S}_1 = 2^4)$	$D_2(\mathcal{S}_2 = 2^5)$	$D_3(\mathcal{S}_3 = 2^6)$	$\#O$
1	Rank Scan	0.989 (0.082)	0.878 (0.305)	0.461 (0.448)	0.000 (0.000)
	RSI($m = 2^3$)	0.950 (0.186)	0.764 (0.370)	0.332 (0.358)	4.305 (5.653)
	RSI($m = 2^5$)	0.998 (0.022)	0.982 (0.098)	0.609 (0.392)	0.520 (0.501)
1.5	Rank Scan	0.922 (0.251)	0.542 (0.455)	0.067 (0.132)	0.000 (0.000)
	RSI($m = 2^3$)	0.843 (0.307)	0.342 (0.354)	0.104 (0.080)	3.920 (2.082)
	RSI($m = 2^5$)	0.983 (0.079)	0.877 (0.236)	0.225 (0.111)	0.055 (0.229)
2	Rank Scan	0.763 (0.410)	0.206 (0.333)	0.043 (0.093)	0.000 (0.000)
	RSI($m = 2^3$)	0.619 (0.382)	0.154 (0.121)	0.089 (0.063)	3.945 (2.385)
	RSI($m = 2^5$)	0.978 (0.090)	0.667 (0.280)	0.208 (0.05)	0.060 (0.238)

After merging the contiguous selected segments, RSI found 30 possible CNVs and the rank scan method selected 34. Figure 3 shows the histograms of the read depths of the selected CNVs. We can see the read depth in the rank scan method is generally larger than that in RSI.

6 Discussion

In this paper we consider a prototypical structured detection setting with the particularity that the null distribution is unknown. When the null distribution is known, various works have shown that a form of scan test achieves the best possible asymptotic power. When the null distribution is unknown, one can alternatively calibrate the scan test by permutation. This has been suggested a number of times in the detection literature. Theorem 1 implies doing this results in no loss of asymptotic power compared to a calibration by Monte Carlo with full knowledge of the null distribution. To circumvent the expense of calibrating by permutation, we propose to scan the

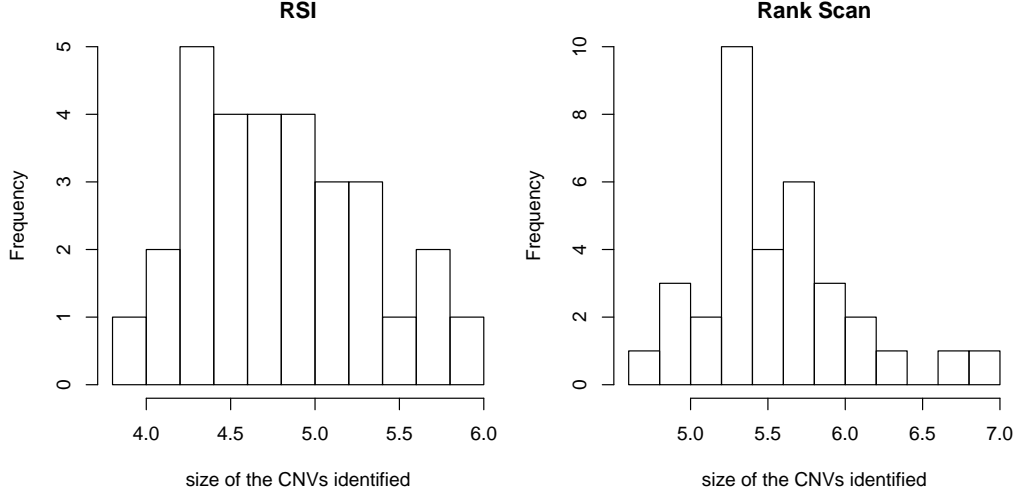


Figure 3: Histogram of the read depths of the selected CNVs in log scale (base 10). Both methods only scan dyadic intervals of lengths from 2^1 to 2^{16} . The RSI used a bin size $m = 400$, while the rank scan was calibrated as for testing.

ranks. Theorem 2 and Proposition 2 imply that this results in very little loss in asymptotic power. In our empirical experiments all three methods perform comparably. Generalizations to multivariate scenarios are also possible (e.g., $X_v \in \mathbb{R}^d$ with $d > 1$). The exact procedure will depend heavily on the specific problem context. For instance, in imaging contexts the entries of X_v correspond to measurements in different wavelengths that might be suitably combined in a single univariate score.

Censoring before permutation. When F_0 is not of compact support, we can enforce it by applying a censoring of the form $\tilde{X}_v = X_v \mathbb{1}_{\{|X_v| \leq t\}} + t \text{sign}(X_v) \mathbb{1}_{\{|X_v| > t\}}$. With a choice of threshold $t = t_N \rightarrow \infty$ slowly (e.g., $t_N = \log \log N$), Theorem 1 applies with $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v$ replaced by $\min_{v \in \mathcal{S}} \theta_v$ and without an upper bound on the θ'_v s. The proof of this result is nearly identical except for very minor modifications. This censoring has the added advantage of making the method more robust to possible outliers.

Other scoring functions. Although rank-sums are intuitive and classically used, any scan based on $h(r_v)$, where h is increasing, is valid. (Recall that r_v is the rank of x_v in the sample.) In two-sample testing, it is known that there is no uniformly better choice of function h . See (Lehmann and Romano, 2005, Sec 6.9) where it is shown that choosing $h(r) = \mathbb{E}(Z_{(r)})$ — where $Z_{(1)} < \dots < Z_{(N)}$ are the order statistics of a standard normal sample — is (in some sense) optimal in the normal location model. Our method of proof applies to a general h .

Unstructured subsets. No permutation approach (including a rank-based approach) has any power for detecting unstructured anomalies. A prototypical example is when \mathcal{S} is the class of all subsets, or all subsets of given size, the latter including the class of singletons.

7 Proofs

7.1 Proof of Theorem 1

Suppose first we are under the null hypothesis. Note that $\mathbf{X} = (X_v, v \in \mathcal{V})$ are IID under the null, and therefore exchangeable. This means that, for any permutation π the marginal distributions

of $\text{SCAN}(\mathbf{X})$ and $\text{SCAN}(\mathbf{X}_\pi)$ are the same. This implies that $\text{SCAN}(\mathbf{X})$ is conditionally uniformly distributed on the set $\{\text{SCAN}(\mathbf{X}_\pi), \pi \in \mathcal{V}!\}$ (with multiplicities) and so

$$\mathbb{P}(|\{\pi \in \mathcal{V}! : \text{SCAN}(\mathbf{X}_\pi) \geq \text{SCAN}(\mathbf{X})\}| \leq \alpha \mathcal{V}!) \leq \frac{\lfloor \alpha |\mathcal{V}! \rfloor}{|\mathcal{V}!|} \leq \alpha,$$

where $\lfloor z \rfloor$ denotes the integer part of z . If there were no ties, the first inequality above would be an equality, but with ties present the test becomes more conservative. For more details on permutation tests the reader is referred to (Lehmann and Romano, 2005).

All that remains to be done is to study the permutation test under the alternative hypothesis. This requires two main steps. First we need to control the randomness in the permutation, conditionally on the observations \mathbf{x} . Once this is done we remove the conditioning.

The key to the first step is the following Bernstein's inequality for sums of variables sampled without replacement from a finite population.

Lemma 2 (Bernstein's inequality for sampling without replacement). *Let (Z_1, \dots, Z_m) be obtained by sampling without replacement from a given a set of real numbers $\{z_1, \dots, z_J\} \subset \mathbb{R}$. Define $z_{\max} = \max_j z_j$, $\bar{z} = \frac{1}{J} \sum_j z_j$, and $\sigma_z^2 = \frac{1}{J} \sum_j (z_j - \bar{z})^2$. Then the sample mean $\bar{Z} = \frac{1}{m} \sum_i Z_i$ satisfies*

$$\mathbb{P}(\bar{Z} \geq \bar{z} + t) \leq \exp \left[- \frac{mt^2}{2\sigma_z^2 + \frac{2}{3}(z_{\max} - \bar{z})t} \right], \quad \forall t \geq 0.$$

This result is a consequence of (Hoeffding, 1963, Th. 4) and Chernoff's bound, from which Bernstein's inequality is derived, as in⁹ (Shorack and Wellner, 1986, p 851). See (Bardenet and Maillard, 2013; Boucheron et al., 2013; Dembo and Zeitouni, 2010) for a discussion of the literature on concentration inequalities for sums of random variables sampled without replacement from a finite set.

Applying this result for a fixed (but arbitrary) set $\mathcal{S}^* \in \mathbb{S}_b$ when π is uniformly drawn from $\mathcal{V}!$ and \mathbf{x} is given, we get

$$\mathbb{P}(Y_{\mathcal{S}^*}(\mathbf{x}_\pi) - \sqrt{|\mathcal{S}^*|} \bar{x} \geq t) \leq \exp \left[- \frac{t^2}{2\sigma_x^2 + \frac{2}{3}(x_{\max} - \bar{x})t/\sqrt{|\mathcal{S}^*|}} \right], \quad \forall t \geq 0,$$

using the same notation as in Lemma 2. Plugging in $t = \text{SCAN}(\mathbf{x})$, noting that $|\mathcal{S}^*| \geq 2^q/(1+2^{-b+2}) \geq 2^q/2$ eventually (because $b \rightarrow \infty$), and using this together with a union bound, we get

$$\mathfrak{P}(\mathbf{x}) \leq |\mathbb{S}_b| \exp \left[- \frac{\text{SCAN}(\mathbf{x})^2}{2\sigma_x^2 + (x_{\max} - \bar{x})2^{-q/2}\text{SCAN}(\mathbf{x})} \right]. \quad (15)$$

(The $\frac{2}{3}$ in the denominator, when multiplied by $\sqrt{2}$, from $|\mathcal{S}| \geq 2^q/2$, is still less than 1.)

Now we proceed by upper bounding the right-hand side of the above inequality by assuming we are under the alternative, which yields an upper bound for the P-value $\mathfrak{P}(\mathbf{X})$. This amounts to controlling the terms $X_{\max} - \bar{X}$, $\sigma_{\mathbf{X}}^2$ and $\text{SCAN}(\mathbf{X})$ under the alternative (upper-case X relates to the random quantities.)

Recall that F_0 has zero mean and unit variance and note that $\mathbb{E}_\theta(X)$ and $\text{Var}_\theta(X)$ are continuous in θ (and thus bounded on the interval $[0, \tilde{\theta}]$).

⁹There is a typo in the statement of the result in (Shorack and Wellner, 1986, p 851), but following the proof one can find the correct result. Where the statement of the result reads $-\frac{\lambda}{2\sigma^2}$ we should have $-\frac{\lambda^2}{2\sigma^2}$ instead

We begin by controlling $\mathbf{X}_{\max} - \bar{\mathbf{X}}$. Let \mathcal{S} denote the anomalous interval under the alternative. We have

$$\bar{\mathbf{X}} = \frac{1}{N} \sum_{v \in \mathcal{V}} \mathbb{E}(X_v) + \frac{1}{N} \sum_{v \in \mathcal{V}} (X_v - \mathbb{E}(X_v)) = O(|\mathcal{S}|/N) + o_P(1) = o_P(1) ,$$

as $N \rightarrow \infty$, since $|\mathcal{S}| = o(N)$, $\theta_v \leq \tilde{\theta}$ for all $v \in \mathcal{V}$, and using Chebyshev's inequality in the second equality. Furthermore, let $\mathbf{X}_{\max, \mathcal{S}} = \max_{v \in \mathcal{S}} X_v$ be the maximum over the anomalous set \mathcal{S} . Let $\bar{\mathcal{S}}$ denote the complement of \mathcal{S} . A union bound together with $\mathbf{X}_{\max} = \mathbf{X}_{\max, \mathcal{S}} \vee \mathbf{X}_{\max, \bar{\mathcal{S}}}$ implies

$$\mathbb{P}(\mathbf{X}_{\max} > x) \leq \mathbb{P}(\mathbf{X}_{\max, \mathcal{S}} > x) + \mathbb{P}(\mathbf{X}_{\max, \bar{\mathcal{S}}} > x) \leq |\mathcal{S}| \bar{F}_{\tilde{\theta}}(x) + |\bar{\mathcal{S}}| \bar{F}_0(x) ,$$

where $\bar{F}_{\theta}(x) = \mathbb{P}_{\theta}(X > x)$ and we used the fact that $\bar{F}_{\theta}(x)$ is monotone increasing in θ - see Section 1.2. For $c \in (0, \theta_{\star} - \tilde{\theta})$, we have

$$\begin{aligned} \bar{F}_{\tilde{\theta}}(x) &= \int_x^{\infty} e^{\tilde{\theta}u - \log \varphi_0(\tilde{\theta})} dF_0(u) \\ &= \frac{1}{\varphi_0(\tilde{\theta})} \int_x^{\infty} e^{-cu} e^{(\tilde{\theta}+c)u} dF_0(u) \leq \frac{\varphi_0(\tilde{\theta}+c)}{\varphi_0(\tilde{\theta})} e^{-cx} . \end{aligned}$$

Using this with the above union bound gives $\mathbb{P}(\mathbf{X}_{\max} > (2/c) \log N) \rightarrow 0$ as $N \rightarrow \infty$. This and the bound on $\bar{\mathbf{X}}$ imply that

$$\mathbb{P}(\mathbf{X}_{\max} - \bar{\mathbf{X}} > (3/c) \log N) \rightarrow 0 .$$

We now consider $\sigma_{\mathbf{X}}^2$. Similarly as before, we have

$$\sigma_{\mathbf{X}}^2 = \frac{1}{N} \sum_{v \in \mathcal{V}} (X_v - \bar{X})^2 \leq \frac{1}{N} \sum_{v \in \mathcal{V}} X_v^2 = \frac{1}{N} \sum_{v \in \mathcal{V}} \mathbb{E}(X_v^2) + \frac{1}{N} \sum_{v \in \mathcal{V}} (X_v^2 - \mathbb{E}(X_v^2)) .$$

On one hand,

$$\begin{aligned} \frac{1}{N} \sum_{v \in \mathcal{V}} \mathbb{E}(X_v^2) &= \frac{1}{N} \sum_{v \notin \mathcal{S}} \text{Var}(X_v) + \frac{1}{N} \sum_{v \in \mathcal{S}} (\text{Var}(X_v) + \mathbb{E}(X_v)^2) \\ &= 1 - \frac{|\mathcal{S}|}{N} + O\left(\frac{|\mathcal{S}|}{N}\right) = 1 + o(1) , \end{aligned}$$

using $\text{Var}(X_v) = 1$ for $v \notin \mathcal{S}$, $\max_{v \in \mathcal{S}} \text{Var}(X_v) < \infty$ and $\max_{v \in \mathcal{S}} \mathbb{E}(X_v) < \infty$ (since $\max_{v \in \mathcal{S}} \theta_v \leq \tilde{\theta}$), as well as our assumption that $|\mathcal{S}| = o(N)$. On the other hand,

$$\frac{1}{N} \sum_{v \in \mathcal{S}} (X_v^2 - \mathbb{E}(X_v^2)) = O_P(1/\sqrt{N}) ,$$

using the fact that $\max_{v \in \mathcal{S}} \mathbb{E}(X_v^4) < \infty$ (since $\max_{v \in \mathcal{S}} \theta_v \leq \tilde{\theta}$) combined with Chebyshev's inequality. We may therefore conclude that

$$\mathbb{P}(\sigma_{\mathbf{X}}^2 \leq 1 + \varepsilon/4) \rightarrow 1 ,$$

with a fixed but arbitrary $\varepsilon > 0$ (we will choose an appropriate value for ε later on).

From Lemma 1 (which does apply to the newly defined \mathbb{S}_b) there is a set $\mathcal{S}^* \in \mathbb{S}_b$ such that $\mathcal{S} \subseteq \mathcal{S}^*$ and $\rho(\mathcal{S}, \mathcal{S}^*) \geq (1 + 2^{-b+2})^{-1/2}$. Note that $\rho(\mathcal{S}, \mathcal{S}^*) = 1 - o(1)$ by the fact that $b \rightarrow \infty$. We then have

$$\begin{aligned} \text{SCAN}(\mathbf{X}) &\geq \mathbf{X}_{\mathcal{S}^*} - \sqrt{|\mathcal{S}^*|} \bar{\mathbf{X}} = \sqrt{|\mathcal{S}^*|} (\bar{\mathbf{X}}_{\mathcal{S}^*} - \bar{\mathbf{X}}) \\ &\geq \sqrt{|\mathcal{S}^*|} \left(\frac{|\mathcal{S}|(N - |\mathcal{S}^*|)}{|\mathcal{S}^*|N} \bar{\mathbf{X}}_{\mathcal{S}} - \frac{N - |\mathcal{S}|}{N} \bar{\mathbf{X}}_{\mathcal{V} \setminus \mathcal{S}} \right) , \end{aligned}$$

where $\bar{\mathbf{X}}_{\mathcal{S}}$ and $\bar{\mathbf{X}}_{\mathcal{V} \setminus \mathcal{S}}$ are the averages of the components of \mathbf{X} over the sets \mathcal{S} and $\mathcal{V} \setminus \mathcal{S}$ respectively. By Chebyshev's inequality,

$$\begin{aligned}\bar{\mathbf{X}}_{\mathcal{S}} &= \frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \mathbb{E}(X_v) + O_P(1/\sqrt{|\mathcal{S}|}) , \\ \bar{\mathbf{X}}_{\mathcal{V} \setminus \mathcal{S}} &= O_P(1/\sqrt{N - |\mathcal{S}|}) .\end{aligned}$$

Recall that we have

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \theta_v \geq \tau \sqrt{\frac{2 \log N}{|\mathcal{S}|}} := \theta_{\dagger} . \quad (16)$$

Note that θ_{\dagger} converges to zero by the assumption on q_l and the fact that τ is fixed. Furthermore $\mathbb{E}_{\theta}(X)$ is increasing in θ (as $\frac{\partial}{\partial \theta} \mathbb{E}_{\theta}(X) = \mathbb{E}_{\theta}(X^2) \geq 0$) and $\mathbb{E}_{\theta}(X) = \theta + O(\theta^2)$ when $\theta \rightarrow 0$ (this can be checked by noting $\mathbb{E}_{\theta}(X) = \int x e^{\theta x} dF_0(x)$ and writing the Taylor expansion of $e^{\theta x}$ around zero). Thus $\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} \mathbb{E}(X_v) \geq \mathbb{E}_{\theta_{\dagger}}(X) = \theta_{\dagger} + O(\theta_{\dagger}^2)$ because $\theta_{\dagger} \rightarrow 0$. Using $\sqrt{|\mathcal{S}^*|} = (1 + o(1))\sqrt{|\mathcal{S}|}$ and $|\mathcal{S}| = o(N)$ we get

$$\text{SCAN}(\mathbf{X}) \geq (1 + o(1))\tau\sqrt{2 \log N} + O_P(1) ,$$

therefore

$$\text{SCAN}(\mathbf{X}) \geq \sqrt{2(1 + \varepsilon/2) \log N} ,$$

with probability tending to one as $N \rightarrow \infty$, where we take ε so that $\tau = \sqrt{1 + \varepsilon}$.

We are ready to make use of the upper bound on the P-value given by (15) and using the condition on q_l we get

$$\begin{aligned}\log \mathfrak{P}(\mathbf{X}) &\leq \log |\mathbb{S}_b| - \frac{2(1 + \varepsilon/2) \log N}{2(1 + \varepsilon/4) + (3/c)(\log N)\sqrt{2^{-q_l+1}(1 + \varepsilon/2) \log N}} \\ &\leq \log |\mathbb{S}_b| - \frac{(1 + \varepsilon/2) \log N}{1 + \varepsilon/4 + o(1)} ,\end{aligned}$$

with probability going to 1. For the size of the approximating net we have

$$\log |\mathbb{S}_b| \leq \log(N4^{b+1}) = \log N + (b+1) \log 4 = (1 + o(1)) \log N , \quad (17)$$

by our assumption on b . Combining these allows us to conclude that $\log \mathfrak{P}(\mathbf{X}) \rightarrow -\infty$ (meaning $\mathfrak{P}(\mathbf{X}) \rightarrow 0$) with probability tending to one, implying that the test has power tending to 1 as $N \rightarrow \infty$, concluding the proof.

7.2 Proof of Theorem 2

The arguments used for the general permutation test apply verbatim under the null hypothesis, so all that remains to be done is to study the performance of the rank scan test under the alternative.

We may directly apply (15), to obtain

$$\mathfrak{P}(\mathbf{r}) \leq |\mathbb{S}_b| \exp \left(- \frac{\text{SCAN}(\mathbf{r})^2}{\frac{N^2}{6} + \frac{N}{2} 2^{-q_l/2} \text{SCAN}(\mathbf{r})} \right) , \quad (18)$$

where we used $\sigma_r^2 = (N^2 - 1)/12 < N^2/12$, $r_{\max} = N$ and $\bar{r} = (N + 1)/2$, so that $r_{\max} - \bar{r} < N/2$. The previous bounds can be directly computed when there are no ties in the ranks, and it is easy to verify that they also hold if ties are dealt with in any of the classical ways (assigning the average

rank, randomly breaking ties, etc). As before, this is a result conditional on the observations $\mathbf{X} = \mathbf{x}$ and hence the ranks $\mathbf{R} = \mathbf{r}$. The next step is to remove this conditioning, which now amounts to controlling the term $\text{SCAN}(\mathbf{R})$.

Let \mathcal{S} denote the anomalous interval under the alternative. From Lemma 1 there is a set $\mathcal{S}^* \in \mathbb{S}_b$ such that $\mathcal{S} \subseteq \mathcal{S}^*$ and $\rho(\mathcal{S}, \mathcal{S}^*) \geq (1 + 2^{-b+2})^{-1/2}$, therefore $\rho(\mathcal{S}, \mathcal{S}^*) = 1 - o(1)$ by the fact that $b \rightarrow \infty$. Since

$$\text{SCAN}(\mathbf{R}) \geq Y_{\mathcal{S}^*}(\mathbf{R}) - \sqrt{|\mathcal{S}^*| \frac{N+1}{2}} ,$$

we focus on obtaining a lower bound on $Y_{\mathcal{S}^*}(\mathbf{R})$ that applies with high probability.

Note that

$$\mathbb{E}(Y_{\mathcal{S}^*}(\mathbf{R})) = \frac{1}{\sqrt{|\mathcal{S}^*|}} \sum_{v \in \mathcal{S}^*} \mathbb{E}(R_v) ,$$

and

$$\text{Var}(\mathbf{R}_{\mathcal{S}^*}) = \frac{1}{|\mathcal{S}^*|} \left(\sum_{v \in \mathcal{S}^*} \text{Var}(R_v) + \sum_{v, w \in \mathcal{S}^*, v \neq w} \text{Cov}(R_v, R_w) \right) .$$

In an analogous fashion to that in (Hettmansperger, 1984), we can make the following claims about the first two moments of the ranks.

Lemma 3. *Suppose $Z_i \sim F_i, i \in [s]$ and independent, also independent of $\{Z_i\}_{i \in [s+1, n]}$ which are i.i.d. and distributed as F_0 . Let R_i denote the rank (in increasing order) of Z_i in the combined sample, and suppose ties are broken randomly. Define*

$$p_{i,j} = \mathbb{P}(X > Y) + \frac{1}{2} \mathbb{P}(X = Y) ,$$

where $X \sim F_i, Y \sim F_j$ are independent. For $i \in [s]$

$$\mathbb{E}(R_i) = \begin{cases} (n-s)p_{i,0} + \sum_{j \in [s], j \neq i} p_{i,j} + 1 & , \text{ when } i \in [s], \\ \frac{n+s+1}{2} - \sum_{j \in [s]} p_{j,0} & , \text{ when } i \notin [s]. \end{cases}$$

Furthermore, as $n, s \rightarrow \infty, s = o(n)$, for $i \in [s]$

$$\text{Var}(R_i) = (\lambda_i - p_{i,0}^2)n^2 + O(sn) ,$$

where

$$\lambda_i = \mathbb{P}(\{X > Y_1\} \cap \{X > Y_2\}) + \mathbb{P}(X = Y_1 > Y_2) + \frac{1}{3} \mathbb{P}(X = Y_1 = Y_2) ,$$

where $X \sim F_i$ and $Y_1, Y_2 \sim F_0$ are jointly independent. Finally, for any $i, j \in [n]$

$$\text{Cov}(R_i, R_j) = O(n) .$$

For the sake of completeness we sketch a proof of Lemma 3 in Appendix 7.5.1. Recall the definition of p_v in (11) and $p_{v,w}$ in Lemma 3. Using the fact that for any i, j we have $p_{i,j} + p_{j,i} = 1$

we get

$$\begin{aligned}
\sqrt{|\mathcal{S}^*|} \mathbb{E}(Y_{\mathcal{S}^*}(\mathbf{R})) &= \sum_{v \in \mathcal{S}^*} \mathbb{E}(R_v) = \sum_{v \in \mathcal{S}} \mathbb{E}(R_v) + \sum_{v \in \mathcal{S}^* \setminus \mathcal{S}} \mathbb{E}(R_v) \\
&= \sum_{v \in \mathcal{S}} \left((N - |\mathcal{S}|)p_v + \sum_{w \in \mathcal{S}, w \neq v} p_{v,w} + 1 \right) + \sum_{v \in \mathcal{S}^* \setminus \mathcal{S}} \left(\frac{1}{2}(N + |\mathcal{S}| + 1) + \sum_{w \in \mathcal{S}} p_w \right) \\
&= |\mathcal{S}|(N - |\mathcal{S}|)\bar{p}_{\mathcal{S}} + \sum_{v \in \mathcal{S}} \sum_{w \in \mathcal{S}, w \neq v} p_{v,w} + |\mathcal{S}| + |\mathcal{S}^* \setminus \mathcal{S}| \frac{1}{2}(N + |\mathcal{S}| + 1) - |\mathcal{S}^* \setminus \mathcal{S}| |\mathcal{S}| \bar{p}_{\mathcal{S}} \\
&= |\mathcal{S}|(N - |\mathcal{S}| - |\mathcal{S}^* \setminus \mathcal{S}|)\bar{p}_{\mathcal{S}} + \frac{1}{2}|\mathcal{S}|(|\mathcal{S}| + |\mathcal{S}^* \setminus \mathcal{S}|) + \frac{1}{2}|\mathcal{S}| + |\mathcal{S}^* \setminus \mathcal{S}| \frac{N+1}{2} \\
&= |\mathcal{S}|(N - |\mathcal{S}| - |\mathcal{S}^* \setminus \mathcal{S}|)(\bar{p}_{\mathcal{S}} - 1/2) + |\mathcal{S}| \frac{N+1}{2} + |\mathcal{S}^* \setminus \mathcal{S}| \frac{N+1}{2} \\
&= |\mathcal{S}|(N - |\mathcal{S}| - |\mathcal{S}^* \setminus \mathcal{S}|)(\bar{p}_{\mathcal{S}} - 1/2) + |\mathcal{S}^*| \frac{N+1}{2} ,
\end{aligned}$$

where $\bar{p}_{\mathcal{S}} = \frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_v$ is the average of p_v over the anomalous set.

Note that for any $v \in [N]$ we trivially have $\text{Var}(R_v) \leq N^2$, and by Lemma 3, $\text{Cov}(R_v, R_w) = O(N)$, so $\text{Var}(Y_{\mathcal{S}^*}(\mathbf{R})) = O(N^2)$. Hence, using Chebyshev's inequality we obtain

$$\begin{aligned}
Y_{\mathcal{S}^*}(\mathbf{R}) - \sqrt{|\mathcal{S}^*|} \frac{N+1}{2} &= \frac{|\mathcal{S}|}{\sqrt{|\mathcal{S}^*|}} (N - |\mathcal{S}| - |\mathcal{S}^* \setminus \mathcal{S}|)(\bar{p}_{\mathcal{S}} - 1/2) + O_P(N) \\
&\geq \rho(\mathcal{S}, \mathcal{S}^*)(N - o(N))\tau\sqrt{2\log N} + O_P(N) ,
\end{aligned} \tag{19}$$

where we used the condition on q_u to conclude that $|\mathcal{S}^*| + |\mathcal{S}^* \setminus \mathcal{S}| = o(N)$. In summary we have

$$\text{SCAN}(\mathbf{R}) \geq c \frac{N}{2\sqrt{3}} \sqrt{2\log N} ,$$

with probability going to 1 as $N \rightarrow \infty$, where $c \in (1, 2\tau\sqrt{3})$.

Plugging this back into (18) and accounting for the condition on q_l we get

$$\begin{aligned}
\log \mathfrak{P}(\mathbf{R}) &\leq \log |\mathbb{S}_b| - \frac{c^2 \frac{N^2}{6} \log N}{\frac{N^2}{6} + \frac{N^2}{2} \frac{c}{2\sqrt{3}} \sqrt{2^{-q_l+1} \log N}} \\
&\leq \log |\mathbb{S}_b| - \frac{c^2 \log N}{1 + o(1)} ,
\end{aligned}$$

with probability going to 1. Noting that the upper bound on $|\mathbb{S}_b|$ in (17) still holds and that $c > 1$ allows us to conclude that $\log \mathfrak{P}(\mathbf{R}) \rightarrow -\infty$ as $N \rightarrow \infty$, hence the test is asymptotically powerful.

7.3 Proof of Proposition 2

Showing this result amounts to relate $p_v \equiv p_{\theta_v}$ with θ_v . This is conveniently done by a Taylor expansion around zero. For ease of presentation let $\theta \equiv \theta_v$ in what follows. When F_0 is discrete, we have

$$p_{\theta} = \int \left(\bar{F}_{\theta}(x) + \frac{1}{2} f_{\theta}(x) F_0(x) \right) dF_0(x) .$$

We expand the integrand seen as a function of θ around $\theta = 0$ up to a second order error term. We have

$$\left. \frac{\partial}{\partial \theta} f_{\theta}(x) \right|_{\theta=0} = x, \quad \left. \frac{\partial}{\partial \theta} \bar{F}_{\theta}(x) \right|_{\theta=0} = \int_{(x, \infty)} u \, dF_0(u) ,$$

where the second identity comes from differentiating inside the integral defining \bar{F}_θ , justified by dominated convergence. Note that $\frac{\partial^2}{\partial \theta^2} f_\theta(x)$ is integrable w.r.t. F_0 when $\theta \in [0, \theta^*)$ and the same holds for $\frac{\partial^2}{\partial \theta^2} \bar{F}_\theta(x)$ as well. Hence let

$$\begin{aligned} c'_0 &:= \int \sup_{\tilde{\theta} \in [0, \theta]} \frac{\partial^2}{\partial \theta^2} f_\theta(x) \Big|_{\theta=\tilde{\theta}} dF_0(x) < \infty, \text{ and} \\ c_0 &:= \int \sup_{\tilde{\theta} \in [0, \theta]} \frac{\partial^2}{\partial \theta^2} \bar{F}_\theta(x) \Big|_{\theta=\tilde{\theta}} dF_0(x) < \infty. \end{aligned}$$

Therefore

$$\begin{aligned} p_\theta &\geq \int \bar{F}_0(x) + \frac{1}{2} F_0(x) + \theta \left(\int_{(x, \infty)} u dF_0(u) + \frac{1}{2} F_0(x) x \right) dF_0(x) - \frac{\theta^2}{2} (c_0 + c'_0/2) \\ &= p_0 + \theta \left(\mathbb{E}_0(X \mathbb{1}_{\{X > Y\}}) + \frac{1}{2} \mathbb{E}_0(X \mathbb{1}_{\{X=Y\}}) \right) - \frac{\theta^2}{2} (c_0 + c'_0/2) \\ &= \frac{1}{2} + \theta \Upsilon_0 - \frac{\theta^2}{2} (c_0 + c'_0/2). \end{aligned}$$

When F_0 is continuous, we have

$$p_\theta = \int \bar{F}_\theta(x) dF_0(x),$$

and similar calculations lead to

$$p_\theta \geq \frac{1}{2} + \theta \Upsilon_0 - \frac{\theta^2}{2} c_0.$$

In summary, we conclude that $p_\theta \geq \frac{1}{2} + \theta \Upsilon_0 + O(\theta^2)$ as $\theta \rightarrow 0$. In addition, note that p_θ is monotonically increasing in θ , by virtue of the fact that $(F_\theta : \theta \geq 0)$ has monotone likelihood ratio. Therefore,

$$\frac{1}{|\mathcal{S}|} \sum_{v \in \mathcal{S}} p_{\theta_v} \geq \frac{1}{2} + \tau \Upsilon_0 \sqrt{\frac{2 \log N}{|\mathcal{S}|}} + O\left(\frac{2 \log N}{|\mathcal{S}|}\right).$$

Finally, using the above bound in (19) and proceeding in an analogous fashion as in Theorem 2 yields the desired result.

7.4 Proof of Proposition 3

We treat each case separately.

Condition (i). The same arguments hold as before under the null, so again we are left with studying the alternative. To deal with smaller intervals, we need a slightly different concentration inequality than before.

Lemma 4 (Chernoff's inequality for ranks). *In the context of Lemma 2, assume that $z_j = j$ for all j . Then*

$$\mathbb{P}(\bar{Z} \geq \bar{z} + t) \leq \exp(-m \sup_{\lambda \geq 0} \psi(t, \lambda)), \quad \forall t \geq 0,$$

where

$$\psi(t, \lambda) := \lambda t - \log \left(\frac{\sinh(\lambda n/2)}{n \sinh(\lambda/2)} \right).$$

Similarly to Lemma 2 this result is also a consequence of Theorem 4 of Hoeffding (1963) and Chernoff's bound. However, with the assumption on z_j in the lemma above we can directly compute the moment generating function of Z_j after using Chernoff's bound instead of upper bounding it, as is classically done to obtain Bernstein's inequality.

In the present context, this yields

$$\mathfrak{P}(\mathbf{r}) \leq |\mathbb{S}| \exp\left(-k\psi(\text{SCAN}(\mathbf{r})/\sqrt{k}, \lambda)\right), \quad \forall \lambda > 0.$$

Note that $x \leq \sinh(x) \leq e^x/2$ and $|\mathbb{S}| \leq N$, hence

$$\mathfrak{P}(\mathbf{r}) \leq N \exp\left(-\lambda\sqrt{k} \text{SCAN}(\mathbf{r}) + \frac{\lambda k N}{2} - k \log(\lambda N)\right), \quad \forall \lambda > 0. \quad (20)$$

The next step is to remove the conditioning $\mathbf{R} = \mathbf{r}$ and bound $\text{SCAN}(\mathbf{R})$. Recall $\text{SCAN}(\mathbf{R}) \geq Y_{\mathcal{S}}(\mathbf{R}) - \sqrt{k}\frac{N+1}{2}$, where \mathcal{S} is the anomalous interval. As in the proof of Theorem 2 we use Lemma 3 to evaluate the terms $\mathbb{E}(Y_{\mathcal{S}}(\mathbf{R}))$ and $\text{Var}(Y_{\mathcal{S}}(\mathbf{R}))$. We have

$$\mathbb{E}(Y_{\mathcal{S}}(\mathbf{R})) = \sqrt{k}(N-k)(\bar{p}_{\mathcal{S}} - 1/2) + \sqrt{k}\frac{N+1}{2},$$

where we use the shorthand notation $\bar{p}_{\mathcal{S}} = \frac{1}{|\mathbb{S}|} \sum_{v \in \mathcal{S}} p_v$. For the variance term, recalling the definition of λ_v from Lemma 3, we note that $\lambda_v \leq p_v$. Hence

$$\text{Var}(R_v) = (\lambda_v - p_v^2)n^2 + O(kN) \leq p_v(1-p_v)N^2 + O(kN) \leq (1-p_v)N^2 + O(kN).$$

Also using $\text{Cov}(R_v, R_w) = O(N)$, we get

$$\text{Var}(Y_{\mathcal{S}}(\mathbf{R})) \leq (1 - \bar{p}_{\mathcal{S}})N^2 + O(kN).$$

According to our assumption, there exists a sequence $\omega_N \rightarrow \infty$ such that

$$\bar{p}_{\mathcal{S}} \geq 1 - \omega_N^{-1}N^{-2/k}.$$

For reasons that become apparent at the end of the proof, we choose $\omega_N \rightarrow \infty$ not too fast (for instance $\omega_N \leq \log N$ suffices). Using Chebyshev's inequality we get

$$\begin{aligned} & \mathbb{P}\left(Y_{\mathcal{S}}(\mathbf{R}) - \sqrt{k}\frac{N+1}{2} \leq \sqrt{k}(N-k)\left(\frac{1}{2} - \omega_N^{-1/4}N^{-1/k}\right)\right) \\ &= \mathbb{P}\left(Y_{\mathcal{S}}(\mathbf{R}) - \mathbb{E}(Y_{\mathcal{S}}(\mathbf{R})) \leq \sqrt{k}(N-k)\left(1 - \omega_N^{-1/4}N^{-1/k} - \bar{p}_{\mathcal{S}}\right)\right) \\ &\leq \mathbb{P}\left(Y_{\mathcal{S}}(\mathbf{R}) - \mathbb{E}(Y_{\mathcal{S}}(\mathbf{R})) \leq -\sqrt{k}(N-k)\left(\omega_N^{-1/4}N^{-1/k} - \omega_N^{-1}N^{-2/k}\right)\right) \\ &\leq \mathbb{P}\left(|Y_{\mathcal{S}}(\mathbf{R}) - \mathbb{E}(Y_{\mathcal{S}}(\mathbf{R}))| \geq \sqrt{k}(N-k)\left(\omega_N^{-1/4}N^{-1/k} - \omega_N^{-1}N^{-2/k}\right)\right) \\ &\leq \frac{N^2\omega_N^{-1}N^{-2/k} + O(kN)}{k(N-k)^2\left(\omega_N^{-1/4}N^{-1/k} - \omega_N^{-1}N^{-2/k}\right)^2} \leq \frac{4N^2\omega_N^{-1}N^{-2/k} + O(kN)}{k(N-k)^2\omega_N^{-1/2}N^{-2/k}} \rightarrow 0, \end{aligned}$$

where the last inequality follows because $\omega_N^{-1/4}N^{-1/k} - \omega_N^{-1}N^{-2/k} \geq \omega_N^{-1/4}N^{-1/k}/2$ eventually as $N \rightarrow \infty$. Hence,

$$\text{SCAN}(\mathbf{R}) \geq \sqrt{k}(N-k)\left(\frac{1}{2} - \omega_N^{-1/4}N^{-1/k}\right),$$

with probability converging to 1 as $N \rightarrow \infty$. Using this with (20) we get

$$\log \mathfrak{P}(\mathbf{R}) \leq \log N + \frac{\lambda k^2}{2} + \lambda k(N-k)\omega_N^{-1/4}N^{-1/k} - k \log(\lambda N), \quad \forall \lambda > 0,$$

with probability tending to 1. Choosing $\lambda = \omega_N^{1/4} N^{1/k} / N$ we get

$$\log \mathfrak{P}(\mathbf{R}) \leq \frac{\omega_N^{1/4} N^{1/k}}{N} k^2 + \frac{N-k}{N} k - \frac{k}{4} \log \omega_N \rightarrow -\infty ,$$

with probability going to 1, where we used that ω_N grows slowly enough for the first term to vanish.

Condition (ii). We can mimic the arguments above. Suppose $k = c \log N$ with arbitrary $c > 0$ and $\bar{p}_S = 1 - (1-\delta) \exp(-\frac{c+1}{c}) := 1 - (1-\delta) f(c)$ with some $\delta > 0$. As before, using Chebyshev's inequality we can show that

$$\text{SCAN}(\mathbf{R}) \geq \sqrt{k}(N-k) \left(\frac{1}{2} - \left(1 - \frac{\delta}{2} \right) f(c) \right) ,$$

with probability tending to 1 as $N \rightarrow \infty$. Plugging this into (20), choosing $\lambda = 1/(Nf(c))$ we get

$$\log \mathfrak{P}(\mathbf{R}) \leq \log N + \frac{k^2}{2f(c)N} + \frac{k(N-k)(1-\frac{\delta}{2})}{N} - k \log f(c) ,$$

with probability going to 1 as $N \rightarrow \infty$. Plugging in $k = c \log n$ and $f(c) = \exp(-\frac{c+1}{c})$ we see that the log of the p -value goes to $-\infty$, which is what we wanted to show.

7.5 Additional results

7.5.1 Sketch proof of Lemma 3

First, assume that there are no ties in the ranks, with probability one. Note that we can write

$$R_i = 1 + \sum_{j \in [n], j \neq i} \mathbb{1}_{\{Z_i > Z_j\}} = 1 + \sum_{j \in [s], j \neq i} \mathbb{1}_{\{Z_i > Z_j\}} + \sum_{j \notin [s], j \neq i} \mathbb{1}_{\{Z_i > Z_j\}} .$$

Taking expectation yields

$$\mathbb{E}(R_i) = \begin{cases} 1 + (n-s)p_{i,0} + \sum_{j \in [s], j \neq i} p_{i,j} & , \text{ when } i \in [s], \\ \frac{n+s+1}{2} - \sum_{j \in [s]} p_{j,0} & , \text{ when } i \notin [s]. \end{cases}$$

since $\mathbb{P}(Z_i = Z_j) = 0$ for $i \neq j$ when there are no ties. The variance and covariance terms can be worked out using the same representation of the ranks as above, but we omit these straightforward computations for the sake of space.

In case of ties, to keep the presentation simple, assume that the distributions of $\{Z_i\}_{i \in [n]}$ are supported on \mathbb{Z} . Then randomly breaking ties in the ranks amounts to using the following procedure. Let $\{\epsilon_i\}_{i \in [n]}$ be independent and uniformly distributed on $(-c, c)$ with $c \leq 1/2$, also independent from $\{Z_i\}_{i \in [n]}$. Consider $Z'_i = Z_i + \epsilon_i$, $i \in [n]$ and let R'_i be the rank of Z'_i in the combined sample $\{Z'_i\}_{i \in [n]}$. Then the joint distribution of $\{R'_i\}_{i \in [n]}$ is the same as that of $\{R_i\}_{i \in [n]}$ when ties are broken randomly.

For instance, for $i \notin [s]$

$$\begin{aligned} \mathbb{E}(R'_i) &= \frac{n+s+1}{2} - \sum_{j \in [s]} \mathbb{P}(Z'_i > Z'_j) \\ &= \frac{n+s+1}{2} - \sum_{j \in [s]} (\mathbb{P}(Z_i > Z_j) + \mathbb{P}(\epsilon_i > \epsilon_j | Z_i = Z_j) \mathbb{P}(Z_i = Z_j)) \\ &= \frac{n+s+1}{2} - \sum_{j \in [s]} p_{j,0} . \end{aligned}$$

The rest of the claims can be worked out similarly.

Finally, when Z_i have arbitrary distributions a similar method can be applied, although it requires a bit more care and one needs to take c approaching zero.

7.5.2 Derivation of Υ_0 in the normal location model

Assume the normal model where $F_\theta = \mathcal{N}(\theta, 1)$. For this case we can simply compute Υ_0 . Since there are no ties with probability 1, we have

$$\Upsilon_0 = \mathbb{E}(X \mathbb{1}_{\{X > Y\}}) = \int_{-\infty}^{\infty} \int_x^{\infty} u f_0(u) du f_0(x) dx.$$

Considering the inner integral we have

$$\int_x^{\infty} u f_0(u) du = \frac{1}{\sqrt{2\pi}} \int_x^{\infty} u e^{-u^2/2} du = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = f_0(x) .$$

Hence

$$\Upsilon_0 = \int_{-\infty}^{\infty} f_0(x) = \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-x^2} dx = \frac{1}{2\sqrt{\pi}} .$$

Therefore we conclude that $1/(2\sqrt{3}\Upsilon_0) = \sqrt{\pi/3}$.

Acknowledgments

We would like to thank Daniel Neill for references on fast scans; Martin Kulldorff for references on permutation scans, for bringing the reference (Jung and Cho, 2015) to our attention, and for helpful comments; Bing Zhou for help with the real data application of Section 5.4; and Tengyao Wang and Richard Samworth for discussions about the properties of Υ_0 . Comments of an anonymous referee helped improve the narrative and presentation. The basic concept of this paper arose out of conversations between some of the authors at the 2013 Mathematical Statistics conference in Luminy, France. This work was partially supported by a grant from the US National Science Foundation (DMS 1223137) and a grant from the *Nederlandse organisatie voor Wetenschappelijk Onderzoek* (NWO 613.001.114).

References

- Addario-Berry, L., N. Broutin, L. Devroye, G. Lugosi, et al. (2010). On combinatorial testing problems. *The Annals of Statistics* 38(5), 3063–3092.
- Arias-Castro, E., E. J. Candès, and A. Durand (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics* 39(1), 278–304.
- Arias-Castro, E., E. J. Candès, H. Helgason, and O. Zeitouni (2008). Searching for a trail of evidence in a maze. *Ann. Statist.* 36(4), 1726–1757.
- Arias-Castro, E., D. Donoho, and X. Huo (2005). Near-optimal detection of geometric objects by fast multiscale methods. *IEEE Trans. Inform. Theory* 51(7), 2402–2425.
- Arias-Castro, E. and G. R. Grimmett (2013). Cluster detection in networks using percolation. *Bernoulli* 19(2), 676–719.
- Balakrishnan, N. and M. V. Koutras (2002). *Runs and Scans with Applications*. Wiley.
- Bardenet, R. and O.-A. Maillard (2013). Concentration inequalities for sampling without replacement. *arXiv preprint arXiv:1309.4029*.

- Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.
- Boutsikas, M. V. and M. V. Koutras (2006). On the asymptotic distribution of the discrete scan statistic. *J. Appl. Probab.* 43(4), 1137–1154.
- Cai, T. T., J. X. Jeng, and H. Li (2012). Robust detection and identification of sparse segments in ultrahigh dimensional data analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(5), 773–797.
- Cai, T. T. and M. Yuan (2014). Rate-optimal detection of very short signal segments. *arXiv preprint arXiv:1407.2812*.
- Cheung, Y. T. D., M. J. Spittal, M. K. Williamson, S. J. Tung, and J. Pirkis (2013, 01). Application of scan statistics to detect suicide clusters in australia. *PLoS ONE* 8(1), e54168.
- Dembo, A. and O. Zeitouni (2010). *Large deviations techniques and applications*, Volume 38 of *Stochastic Modelling and Applied Probability*. Berlin: Springer-Verlag. Corrected reprint of the second (1998) edition.
- Desolneux, A., L. Moisan, and J.-M. Morel (2003). Maximal meaningful events and applications to image analysis. *Ann. Statist.* 31(6), 1822–1851.
- Ernst, J., P. Kheradpour, T. S. Mikkelsen, N. Shores, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* 473(7345), 43–49.
- Flenner, A. and G. Hoyer (2011). A Helmholtz principle approach to parameter-free change detection and coherent motion using exchangeable random variables. *SIAM J. Imaging Sci.* 4(1), 243–276.
- Guerriero, M., P. Willett, and J. Glaz (2009, July). Distributed target detection in sensor networks using scan statistics. *Signal Processing, IEEE Transactions on* 57(7), 2629–2639.
- Hall, P. and J. Jin (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.* 38(3), 1686–1732.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58, 13–30.
- Huang, L., M. Kulldorff, and D. Gregorio (2007). A spatial scan statistic for survival data. *Bio-metrics* 63(1), 109–118.
- Jeng, X. J., T. T. Cai, and H. Li (2010). Optimal sparse segment identification with application in copy number variation analysis. *Journal of the American Statistical Association* 105(491), 1156–1166.
- Jiang, T. (2002). Maxima of partial sums indexed by geometrical structures. *Ann. Probab.* 30(4), 1854–1892.
- Jung, I. and H. Cho (2015). A nonparametric spatial scan statistic for continuous data. *International Journal of Health Geographics* 14(1), 30.
- Kabluchko, Z. (2011). Extremes of the standardized gaussian noise. *Stochastic Processes and their Applications* 121(3), 515–533.
- Kulldorff, M. (1997). A spatial scan statistic. *Comm. Statist. Theory Methods* 26(6), 1481–1496.
- Kulldorff, M., R. Heffernan, J. Hartman, R. Assuncao, and F. Mostashari (2005). A space-time permutation scan statistic for disease outbreak detection. *PLOS Medicine* 2(3), 216.
- Kulldorff, M., L. Huang, and K. Konty (2009). A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics* 8(1), 58.
- Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer

- Texts in Statistics. New York: Springer.
- McFowland, E., S. Speakman, and D. B. Neill (2013). Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research* 14(1), 1533–1561.
- Neill, D. B. (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74(2), 337–360.
- Neill, D. B. and A. W. Moore (2004). Rapid detection of significant spatial clusters. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 256–265. ACM.
- Nichols, T. E. and A. P. Holmes (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping* 15(1), 1–25.
- Perone Pacifico, M., C. Genovese, I. Verdinelli, and L. Wasserman (2004). False discovery control for random fields. *J. Amer. Statist. Assoc.* 99(468), 1002–1014.
- Sharpnack, J. and E. Arias-Castro (2014). Exact asymptotics for the scan statistic and fast alternatives. *arXiv preprint arXiv:1409.7127*.
- Sharpnack, J. and A. Singh (2010). Identifying graph-structured activation patterns in networks. In *Advances in Neural Information Processing Systems*, pp. 2137–2145.
- Sharpnack, J. L., A. Krishnamurthy, and A. Singh (2013). Near-optimal anomaly detection in graphs using lovász extended scan statistic. In *Advances in Neural Information Processing Systems*, pp. 1959–1967.
- Shorack, G. R. and J. A. Wellner (1986). *Empirical processes with applications to statistics*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. New York: John Wiley & Sons Inc.
- Siegmund, D. and E. S. Venkatraman (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *Ann. Statist.* 23(1), 255–271.
- Wallenstein, S. (2009). Joseph naus: Father of the scan statistic. In *Scan Statistics*, pp. 1–25. Springer.
- Walther, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *The Annals of Statistics* 38(2), 1010–1033.
- Zhao, M. and V. Saligrama (2009). Anomaly detection with score functions based on nearest neighbor graphs. In *Advances in Neural Information Processing Systems*, pp. 2250–2258.